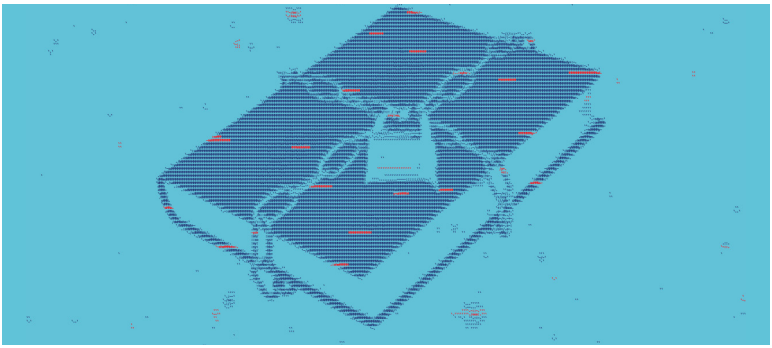
Platform Accountability Through Digital "Poison Cabinets"

DATA AND DEMOCRACY



Platform Accountability Through Digital "Poison Cabinets"

By John Bowers, Elaine Sedenberg, and Jonathan Zittrain

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In October 2020, the Knight First Amendment Institute at Columbia University convened a virtual symposium, titled "Data and Democracy," to investigate how technological advances relating to the collection, analysis, and manipulation of data are affecting democratic processes, and how the law must adapt to ensure the conditions for self-government. This symposium was organized by the Institute's 2019-2020 Senior Visiting Research Scholar, Yale Law Professor Amy Kapczynski, and co-sponsored by the Law and Political Economy Project at Yale Law School.

The essays in this series were originally presented and discussed at this two-day event. Written by scholars and experts in law, computer science, information studies, political science, and other disciplines, the essays focus on three areas that are both central to democratic governance and directly affected by advancing technologies and ever-increasing data collection: 1) public opinion formation and access to information; 2) the formation and exercise of public power; and 3) the political economy of data.

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, Executive Director; Katy Glenn Bass, Research Director; Amy Kapczynski, Senior Visiting Research Scholar; Alex Abdo, Litigation Director; and Larry Siems, Chief of Staff. The essay series was edited by Glenn Bass with additional support from Lorraine Kenny, Communications Director; A. Adam Glenn, Writer/Editor; and Madeline Wood, Communications and Research Coordinator.

The full series is available at knightcolumbia.org/research/

INTRODUCTION

AS DENAZIFICATION AND RECONSTRUCTION efforts ramped up across Germany in the wake of World War II, Germans and occupying Allied forces found themselves struggling with difficult questions around censorship and preservation, among them what to do with decades worth of Nazi writings. While books including *Mein Kampf* were banned for a period of years after the war as part of an effort to redesign Germany's political system, reformulate its national identity, and reverse years of indoctrination,[1] the prospect of purging them completely—leaving no copy unpulped[2]—carried unwelcome echoes of Nazi campaigns of book burning and repression. With collective memory and public remembrance as key pillars of the country's transitional process, the total erasure of past horrors would have been sorely out of step.[3]

To balance these two competing needs—on one hand, to limit the circulation of ugly, potentially corruptive materials; on the other, to preserve them as objects of study and reflection—Germany's reformers drew inspiration from an archival institution dating back centuries: the Giftschrank.[4] Giftschränke (literally "poison cabinets") are cordoned-off sections of libraries, whether wings, rooms, or designated shelves and cabinets, built to house

materials deemed unfit for widespread circulation.[5] Rather than destroying such materials, German censors have, at various times in history, elected to lock them away in Giftschränke, leaving them accessible only to accredited researchers.[6] Put simply, Giftschränke enable access management without deletion—a means of restricting the availability of harmful materials without purging them from the historical record. Where full deletion places artifacts decisively outside of the grasp of those who might otherwise learn from them, the Giftschrank does not.

In keeping with this ethos, copyright over Hitler's writings was transferred to the Bavarian state government, which—without literally locking all of the texts in a vault—prohibited reprints, and established controls for where and how copies could be held and accessed.[7] The copyright expired at the end of 2015, and reprints have since been published.[8]

The challenges confronted by archivists in postwar Germany—and the solution offered by Giftschränke—reflect a broader issue common to many areas of archival practice. Art, writing, and expression that captures or even champions toxic ideas can be essential to understanding the development of society, ideology, and politics. Had all Nazi writings been purged in the wake of WWII, our historical record would be deeply impoverished. Had all written documentation of the hateful ideas and assumptions of the Confederacy been burned in the wake of the Civil War, the history of slavery and race would stand even less complete.[9]

But an archivist must take extraordinary care in the act of committing such ugliness and toxicity to the historical record—of forever giving it a place in history books, and even making it available to those who might be seduced by it. Archival choices made around *what* content is preserved in the historical record determine not only whose voices and stories are heard, but also what evidence exists to understand and definitively document atrocities.[10] Preservation and the facilitation of remembrance carry weighty responsibilities to the public at large. Indeed, when Hitler's *Mein Kampf* was offered for sale in Germany in 2016 for the first time since WWII, it came in the form of a 2,000-page critical edition, with the clunky hatefulness of the original text juxtaposed with meticulously researched rebuttal.[11] And more prosaic concerns also apply to the prospect of limiting access to some materials, from the privacy of those referenced in archival materials to legal sensitivities arising out of data protection laws.

In recent years, the ethical character of archival preservation has found new relevance in yet-unsolved questions around the frequent nonpublic censorship of speech by private online platforms. Here, the Giftschrank may serve as a model to consider and balance competing values and challenges.

With the ascendance of social media as a primary gateway for public expression has come greater awareness of the concrete harms implicated by "lawful but awful" content like hate speech and disinformation.[12] As platforms have sought to delimit the bounds of permissible speech more restrictively than national laws require (or than the First Amendment would permit of a public actor), they have found themselves taking on the censor's role. Their task is a challenging one: to set irreducibly normative rules and standards for what speech they will allow to find, or will affirmatively recommend to, an audience; to enforce those rules as consistently as possible across cultural and political lines; and to explain their efforts to a rightly skeptical public. Content governance admits of no closed-form, permanent solution, given the instability of trends in and definitions of harmful content, as well as continual tactical adaptation on the part of those who willfully proliferate such content.[13]

As the platforms have increasingly undertaken this momentous task,[14] they've largely refrained from disclosing the specific content moderation actions they've taken. Nor have they consistently maintained researcher-accessible records of the content subject to removal. Across and within platforms, removed content is handled in a wide variety of ways in accordance with legal and technical needs. Sometimes content is retained on a temporary basis to support appeals processes[15] or comply with governmentally imposed data retention mandates.[16] Sometimes it is preserved and even shared in accordance with applicable law as part of internal research and transparency efforts.[17] And sometimes it is deleted entirely. The lack of accessible, well-documented indexing and archival schemes for platform-moderation decisions represent a threat to the accountability of one of the largest speech-governance projects ever undertaken; to the efficacy of that project in limiting speech-related harms while championing freedom of expression; and to the historical record. This is an industry-wide, contemporary problem that spans platforms of different shapes, sizes, and digital cultures.

This lack of organized documentation is not entirely without its

reasons—including the archival quandaries involved in committing ugly or outright harmful content to a permanent record. Long-term preservation of material for research purposes remains a challenge for private companies in general throughout recent history, let alone for delicate policy decisions. While we might not worry about the corruptive power of objectionable content in quite the same manner as a 16th century Bavarian censor might, or a fragile post-fascist transitional government, the data that would make up an archive of content moderation activity is nonetheless sensitive. Platforms are rightly wary of aggregating repositories of harmful content for fear of increasing that content's visibility, raising privacy concerns, and inviting legal exposure.

The Giftschrank model stands to be enormously useful in navigating these complexities, and establishing a framework for transparency, accountability, and historical preservation on the part of online platforms. The conversation around platforms' documentation of content moderation practices has so far focused largely on real-time or one-off[18] disclosure measures—like application programming interfaces (APIs) for researchers,[19] or data dumps[20]—which make the sorts of concerns above particularly salient. Less well-explored have been the prospects of longer-term, more archivally minded approaches, those that focus less on up-to-the minute telemetry, and more on establishing a longer-term record of how content moderation practices have developed and been implemented over time. While not a substitute for more real-time data sharing, such an approach—complete with the curation, redaction, and access limitation measures that frequently accompany sensitive archival projects—could put those studying content moderation in broader perspective on much firmer footing, and furnish new ground truth for content moderation conversations.

Take, for example, the prospect of a long-term archive of the unprecedented wave of misinformation and disinformation related to the 2020 COVID-19 pandemic. Platforms are reasonably hesitant to build and release, even to accredited researchers, compendia of the harmful narratives that they've made a moderation priority.[21] Providing reasonably complete information about those narratives would mean offering context around engagement and authorship, potentially implicating risks to user privacy. It also would mean giving information about how and why the content was removed, inviting

criticism of its handling of an ongoing crisis while potentially exposing sensitive tactical information to adversaries. And if researchers are to have access to the underlying content itself—which would surely be helpful in many cases—data sharing could create a risk of recirculation, potentially undermining the purpose of moderation.

If the platforms were to delay the release of such an archive for months or years, many of these challenges would become more manageable through curation (archivists will have more time to redact and engineer data for privacy), and reduced sensitivity (COVID-19 misinformation won't be as dangerous once a critical mass of the population has been vaccinated). And while the immediate strategic relevance of the data itself would clearly decline over such an interval, it would still represent an absolutely essential resource for researchers wanting to better understand how platforms manage disinformation in times of crisis, or how the specific dynamics of a pandemic reoriented enforcement priorities. Put simply, a long-term archival approach could take advantage of the fact that the risks posed by the sharing of platform data on COVID-19 misinformation and disinformation are likely to fade much faster than that data's usefulness to researchers.

A Giftschrank could provide the foundation for an archival project of this kind envisioned and carried out by the platforms, ideally one supported by the expertise, coordination, and oversight of seasoned archival institutions. In the process of removing or submerging content that violates a platform's content moderation standards, platforms might create records both of the action taken, the reasons for it, and of the content in question (potentially to include a full copy). Rather than releasing them immediately, these records would then be placed in a secure archive—a Giftschrank—redacted and minimized as necessary to account for privacy concerns, and eventually made available in a controlled fashion to accredited researchers studying harmful content and related moderation practices. By relying on the independent institutions, norms, and standards of archival science and practice, devoted as they are to the responsible provision of access to sensitive materials,[22] the platforms would be tapping into centuries of collected expertise around the responsible management of information. In doing so, they would demonstrate a new degree of respect for the gravity of the role in which they find themselves.

Platforms should not—and likely cannot—go it alone in taking this archival turn. Rather, the design, administration, and oversight of a Giftschrank could be coordinated by independent civil society organizations with a clear public purpose. Indeed, archival expertise is already concentrated among such organizations, including academic libraries and archives. The involvement of these less commercially oriented experts throughout a Giftschrank system would serve a vital accountability function, tying new platform archiving efforts to a real institutional commitment stretching beyond the walls of a single firm. Government, too, has a role to play, whether by providing the legal protections needed to make promising new archival efforts legally feasible, by developing data sharing and transparency mandates, or by some combination thereof.

Properly designed and executed, a Giftschrank approach applied to online platforms' content moderation challenges and practices could satisfy at least two socially important archival objectives. First, over a timeline of months or years, it could bolster the accountability, usefulness, and legitimacy of platforms' content moderation efforts by enabling researchers to analyze, critique, and ultimately help improve them. Second, over a period of years or decades, it could preserve for future scholars and historians vast amounts of vital primary-source material of intense relevance to contemporary politics and society—material that would otherwise be lost through deletion procedures, impoverishing the historical record.

This paper seeks to evaluate the Giftschrank model's theoretical effectiveness as a documentation and accountability framework for contemporary internet platforms, as well as its feasibility within a corporate setting. Part I further develops the key characteristics of the Giftschrank model and explores the relevance of archival approaches to problems of accountability faced by platforms engaged in content moderation. Part II develops some design considerations for a platform Giftschrank and suggests some probable use cases for the model based on salient areas of content moderation. Part III examines the limitations and implementation risks of the Giftschrank model and proposes measures for mitigating them. Building on that analysis, Part IV suggests some pragmatic starting points for implementation of the Giftschrank concept.

I. AN OLD SOLUTION TO A NEW PROBLEM

THE ADAPTATION OF PRINCIPLES and methods undergirding the Giftschrank model to the contemporary platform context needn't be a tortured exercise in historical analogy. There are plenty of differences, in motivation, scale, and capabilities, between the work of a pre-digital German censor and that of a global social media platform. Even so, the two exercises in information control each center on a complex balance among transparency, historical preservation, and public health interests.[23] Today's social media platforms have supercharged old-style censorship—through new levels of granularity in control over content, global-scale human review architectures, and automation—without driving and adopting accompanying innovations in archival technology. To mitigate the risks of new content-shaping paradigms, they must bring the two back into equilibrium. Examining the Giftschrank in historical context lends clues as to how they might do so.

A. The Giftschrank

As mentioned at the outset, the post-World War II era was far from the first time that German authorities sought a flexible means of limiting access to materials deemed unfit for public consumption, or even dangerous. For centuries, German libraries, often acting at the behest of government, kept books and pamphlets considered sensitive or dangerous sequestered under lock and key. These materials ranged from the pornographic to the socially and politically subversive. The Giftschrank—as known by that name—began in the 16th century with the Duke of Bavaria, who opted to restrict rather than destroy hundreds of volumes deemed heretical, so as to better enable members of his court and church allies to grasp the contentions being put forth by sectarian enemies.[24] Giftschränke saw continued use through the centuries that followed, often to contain writings deemed to be at odds with the moral and political standards of society.[25]

While unambiguously a means of censorship, the Giftschrank model reflects an effort to apply what we might today think of as a "public health" calculus to the accessibility of information.[26] It captured a sensibility that there were things to be learned, one way or another, from materials deemed objectionable, offensive, or dangerous—whether tactically, in the interest of

"knowing your enemy," or out of a belief that ideas unfit for public consumption can nonetheless be of value or interest in some limited contexts.[27] (For an illustration of this sensibility in contemporary fiction, one need look no further than Harry Potter's "Restricted Section"—a cordoned-off zone of the Hogwarts library sequestering tomes on taboo subdomains of wizardry.[28])

Indeed, libraries actively sought out books with which to fill their Giftschränke. In 1819, Bavaria's State Library purchased—at considerable expense—a trove of more than 2,900 works of erotic literature collected by the statesman Franz von Krenner.[29] These texts made their way into the library's Giftschrank,[30] and later—thanks to their preservation and indexing—became the subject of an 1889 book.[31] Only in 1967 were they fully integrated into the library's public catalog.[32]

With this utility came evolution beyond the simplicity of a locked cabinet—and significant variation across time and space. As the archival profession developed and libraries, publication volume, and readership grew, Giftschrank-like restricted sections became more formalized and established as part of the topology of libraries and archives. Access to materials deemed dangerous or corruptive was entrusted to professional librarians, with restrictions enforced by institutional and technological controls.[33] In the case of the von Krenner collection, for example, two separate keys were needed to open the Giftschrank containing the erotic volumes.[34] (As radio producer Sam Greenspan notes on the podcast "99% Invisible," the system was not dissimilar to those later used to safeguard nuclear launches.)[35]

Indeed, the principles behind the Giftschrank have been deployed even in cases where the physical sequestering of materials was impossible, or undesirable—as with Hitler's *Mein Kampf* and other writings, of which millions of copies were in circulation by the time denazification efforts began.[36] In these cases, libraries have turned to what Stephen Kellner of the Bavarian State Library calls a "virtual Giftschrank" approach—systems of controls and restrictions on the lending process intended to add friction and verification steps to the process of getting one's hands on potentially harmful materials.[37] This very approach has become even more fraught in the internet age, as libraries struggle to determine whether or not to make their previously paper-only collections of Nazi materials, requiring an in-person visit to view, freely available in digital form.[38]

Regardless of the specifics of its form, a Giftschrank is meant to strike a careful balance between two competing interests: the advantages to preserving access to harmful content for those needing to understand or scrutinize it, and the potential harms arising from the preservation and inclusion in a library or archive. But the mechanics of the Giftschrank cannot themselves ensure that such a balance will be successfully struck. Rather, they simply create the opportunity for curation and access control on the part of professional librarians and archivists, professionals bound by norms and expectations meant to align their conduct with the interests of the public. The success or failure of a Giftschrank by any criterion is ultimately dependent on how these professionals use the opportunities for intervention that the Giftschrank affords—when and to whom they offer access to its contents, what they qualify for inclusion and exclusion, and how effectively those controls align with the values they are meant to serve.

Indeed, Giftschränke can just as easily be tools of repression—and expressions of bureaucratic obstructionism—as enablers of responsible information management.[39] Our sense of the balance between the upsides of information access and the costs of exposure to and preservation of harmful content has rightfully shifted over the past four centuries. The Giftschrank may take on a liberal sheen as an alternative to book burning—which indeed it may have been, for a period of centuries—but it feels anachronistic and authoritarian today, given that the large-scale restriction of access to books is anathema in most liberal democracies. It is not particularly surprising that much of the literature around "true" Giftschränke—that is to say, literal holding areas for restricted texts—in the latter half of the 20th century focuses on East Germany, where they were used to store, among other things, materials reflecting Western culture and ideology.[40] For many East German researchers, the sensitive information housed in the German Democratic Republic's Giftschränke remained almost entirely inaccessible—the restrictions served a bureaucracy committed to a unitary narrative of the world.[41]

So, to relegate anything to a Giftschrank, we might reasonably say, would be a decision of enormous weight. In 1644, less than a century after the first Giftschrank was incorporated in Bavaria, John Milton wrote his *Areopagitica*. Taking aim at the recently published Licensing Order of 1643—which established governmental prepublication censorship—*Areopagitica* offers

a classic defense of the public sphere.[42] To the devout Milton, censorship of valuable ideas is a deep and consequential sin, in that "he who destroys a good book, kills reason itself, kills the image of God."[43] But much censorship[44] that might be considered righteous can bring about serious harms of its own. Turning again to religious allegory, Milton declares that "we bring not innocence into the world, we bring impurity much rather; that which purifies us is trial, and trial is by what is contrary."[45] *Areopagitica* and its ideas still hold great currency in our contemporary conversations around the public sphere—Justice Brennan's majority opinion in the landmark First Amendment case *New York Times v. Sullivan* cites *Areopagitica* alongside John Stuart Mill's *On Liberty* in warning against the prospect of harmful self-censorship by the press.[46]

Given that it stands to introduce new opacity and inequities in access to information, the Giftschrank must respect and contend with that centuries-long legacy. To represent an affirmative move towards transparency and accountability, the Giftschrank must enable the preservation of material that it would otherwise be irresponsible to keep around—it must provide an alternative to deletion for which no equally appealing and less restrictive alternative exists.

In the case of contemporary social media platforms, for which the public disclosure of detailed and representative data relating to content moderation faces currently insurmountable legal, strategic, and privacy barriers, this standard may well be fulfilled. The type of material that would be included in a content moderation archive differs fundamentally from the traditional targets of censorship—books, pamphlets, newspapers—for which archivists tend to default to open access. Much of it, however vile, would reflect pronouncements intended for circles of family and friends, made on the part of individuals. Sharing it publicly would run the risk of violating the privacy of users unprepared for public scrutiny, while raising other risks like reamplification. The Giftschrank approach, which carefully restricts the availability of data to a small group of accredited researchers, after best-effort de-identification and other procedures, may offer a narrow path forward for meaningful archiving. The section that follows will explore some of the challenges confronted by today's platforms, and how longer-term archival approaches might hold promise as a means of addressing them. Part II will

consider the Giftschrank model as a pragmatic means of putting that promise into practice.

B. Content moderation and the archival profession

Today's social media companies have assumed the task of moderating a diverse and contentious public sphere across cultural, political, and geographical boundaries. Their approach to this momentous task, generally speaking, has been to develop—and publish, though not in full—sets of rules and standards meant to draw a boundary between acceptable and unacceptable content by nature of what remains visible (or actively amplified) to other users.[47] The sophistication and granularity of these terms of service, and generally speaking, the restrictiveness, which platforms tend to apply on a global basis, has increased markedly over time as platforms have grown in both scale and experience.[48]

Platforms are in a position to impose a wide range of controls on the visibility and circulation of content—takedowns and restrictions of the sort contemplated in many First Amendment cases and hypotheticals are just a few arrows in an overflowing quiver. Short of removing content entirely, platforms can "downrank" or clamp the virality of objectionable content, minimizing its circulation,[49] or label it with user-interface elements providing further context and scrutiny.[50] Platforms can also intervene on a user or group level, via account removals and restrictions or even "shadow bans," which quietly downrank or hide a user's contributions without informing the user. (Such an approach can trick rulebreakers who might otherwise just make a new account instead of shouting into the void.)[51]

The speech not permitted by platforms is typically much more broad than that targeted by laws governing speech—which in turn are limited in their scope by provisions like the First Amendment.[52] Indeed, much of what the platforms censor falls under the rubric of speech protected from government censorship in some jurisdictions, from cruel mockery[53] to disinformation[54] and hate speech.[55] This represents an unprecedented development, placing the speech of millions around the world on the platforms of a select few private companies, giving their policies and practices lasting impact.

With these new powers has come pressure to consider new mechanisms for accountability and transparency aimed at ensuring their responsible

exercise. And while new experiments for content decisions, like Facebook's Oversight Board,[56] reflect platforms' discomfort with the degree of power they currently wield—and a will to reallocate some of that power outside of their own structures[57]—the fact remains that platforms' arsenals of moderation techniques are generally deployed in ways that lack the traditional markers of accountability. Online platforms are free to set more or less whichever rules they see fit, without facing any obligation to provide explanation or the opportunity for remedy. (Indeed, attempts to impose such obligations by governments would themselves contend with free-speech limitations, with the platforms as the speakers.) More important for our purposes, the platforms are not subject to meaningful disclosure requirements in relation to the content they restrict.[58]

Prospects for legislatively imposing such requirements on the platforms themselves are murky. In the U.S., the body of law protecting platforms' autonomy in moderation is robust, supported by cornerstones like the First Amendment and Section 230 of the Communications Decency Act.[59] The political feasibility of proposals for sweeping reform of the latter remains in doubt, not least because of an absence of consensus of the ways in which Section 230 falls short. In the meantime, much of the focus—buttressed in no small part by public pressure—has shifted to questions of corporate social responsibility, examining what sorts of measures can and should be implemented of their own volition. Given that platforms function as for-profit companies, progress on a voluntary basis, both on forming policy and on enforcing it, is dependent on the identification of common interests and incentives.

In the case of measures aimed at promoting accountability around content moderation practices, such alignment runs deep. Platforms are facing down a serious deficit of public trust,[60] criticism from the academic and research communities, and regulatory pressure in the U.S. and around the world.[61] These pressures have arisen not just from a discomfort with platforms' structural dominance over online speech, but also from a well-substantiated sense that they have at times conceived their rules improperly[62] and applied them unevenly.[63] The Facebook Oversight Board and other measures like it, meant to provide accountability and transparency around when and how content decisions are made, and even to expand the groups

of people involved in making them,[64] represent a step towards a new era of content governance oriented first and foremost towards process-driven legitimacy.[65] In other words, rather than venturing iteration after iteration of their rule sets with the hope of getting in step with the expectations of the broader public, the platforms are beginning to look for new ways to engage the public—or at least a limited subset of people outside their own executives and staff—in the process of content policymaking.

But even the most ambitious of these efforts faces a severe constraint. Inclusivity and accountability in the design of rules and standards is only part of the equation. Accountability in the implementation of those guidelines on a platform-wide level is another. Platforms need to be able to demonstrate that the products of content policymaking processes—however accountable those processes may be—are consistently translated into the expected forms of action. Without broad public faith in this tight coupling between policy and practice, public legitimation of content moderation will remain out of reach.[66] However promising the Facebook Oversight Board may appear, for example, its legitimation in the public eye will ultimately depend on the extent to which it is perceived as making well-founded decisions based on a representative and comprehensible deliberative process. Faith in the process likely won't arise from charter documents alone—it will also require the scrutiny of actual data.

To put it more broadly, building public faith in platform governance will mean embracing new transparency measures that enable independent experts to examine platforms' content moderation practices on a comprehensive or highly representative empirical basis. Researchers from outside of the platforms will need to be allowed to take part in the process of examining how content moderation policies are being implemented in practice, and assessing their effectiveness. Independent fact-finders should be able to rigorously explore questions around whether policies are being enforced consistently, how looser standards are being interpreted by front-line content moderators (human or automated), and how moderation practices are maturing over time. The platforms should not be able to get by on anything along the lines of "take our word for it."

And accountability isn't the only advantage of better recordkeeping and data sharing on the part of the platforms. A better-informed research

community would be able to build an empirically grounded field around the study of harmful content and interventions against it.[67] That's not to mention the fact that harmful content is a deeply salient part of our information ecosystem, and that the historical record would be incomplete without thorough documentation of its manifestations. The ability of future scholars and practitioners to draw meaningfully on the lessons of our time will depend in large part on how well we archive.

So, assuming that platforms are turning in earnest towards binding accountability measures, why haven't they focused more extensively on building and sharing archives of their content moderation practices, along with records of the content those moderation practices were targeted against?

It's a challenging question, and one with any number of possible answers. Most immediately obvious among them might be legal provisions like the European Union's "right to be forgotten," which stand to impose new requirements, restrictions, and potentially liabilities on the retention and sharing of user-generated content.[68]

But even more fundamental are questions of privacy and asset toxicity raised by archives of harmful content. Platforms make moderation decisions on the basis of a complex set of factors, many of which go beyond the *content* of content, further considering metadata like the identity, ongoing behavior, and social graphs of the people and organizations disseminating and engaging with it.[69] Low-level review of content moderation decision making is difficult to disentangle from the delicate privacy considerations that arise whenever such information is subjected to analysis and scrutiny. Platforms looking to share data with researchers are therefore put in the difficult position of having to redact and de-identify appropriately without scrubbing essential context.

And even setting these privacy considerations aside, the fact remains that platforms moderate when they believe that a piece of content or pattern of behavior presents a meaningful risk to their services or users. The material and metadata that would wind up in content moderation archives would, by their very nature, capture or point to exactly the sort of speech and behavior that platforms consider anathema. The consequences could be substantial—retention of legally sensitive material, risks of reamplification[70] if sharing conditions are too generous or leaks materialize, and the possibility of a

"trophy" effect whereby bad actors seek inclusion[71]—making the archives toxic assets. Without accepted industry-wide norms around what "good" or "sufficient" accountability looks like—or well-trodden roadmaps of how to implement archiving in support of it—the activation energy required for meaningful progress remains elusive.

So far, platforms have not found satisfying answers to these questions—nor do we offer fully formed answers to them here. But it's worth noting that they're some of the same questions that the archival profession has grappled with for centuries, including with the specific objective of ensuring the accountability of governments, organizations, and individuals.[72] As a paragraph from the Code of Ethics adopted by the Society of American Archivists concisely states,

> Archivists formulate and disseminate access policies that encourage ethical and responsible use. They work with creators, donors, organizations, and communities to ensure that any restrictions applied are appropriate, well-documented, and equitably enforced. When repositories require restrictions to protect confidential and proprietary information, such restrictions should be applied consistently. Archivists should seek to balance the principles of stewardship, access, and respect.[73]

As that same Code of Ethics makes clear in its opening lines,[74] archiving is a profession built on strong norms and best practices, developed over the course of millions of ethical encounters across thousands of professional careers. As platforms seek to "balance the principles of stewardship, access, and respect" in their own navigation of data sharing and archiving practices, they should look to those learnings for guidance.

Among them is the archival profession's focus on long-term custodianship of information—a concept with which online platforms, all of them relatively recent in their origins, have not yet had occasion to fully internalize.[75] Indeed, as mentioned previously, much of the conversation around data sharing in relation to harmful content has focused on the immediate or near term. That focus is not misplaced: The ecosystems around harmful content are often complex and volatile, and understanding harms—and failures to prevent those harms—sooner can mean stopping them before

they metastasize. But there's been less focus on a complementary but distinct approach to expanding insight and accountability around content moderation—the creation of archival structures that enable the analysis, assessment, and advising of content moderation practices over long periods of time. Legitimacy and accountability will not be secured overnight, but will rather require the development of a long-term evidentiary record of a consistent relationship between accountable content policy processes and actual content moderation practices. The Facebook Oversight Board's rulings, for example, will be so much paper if their implementation cannot be verified. Longitudinal data is indispensable, even if it does not reveal the exact contours of present reality.

Long-term archival approaches to accountability in content moderation may mitigate some of the problems encountered by shorter-term forms of data sharing—including those of privacy and asset toxicity contemplated above. For one, longer-term approaches leave more time for curation—the redaction or de-identification of highly illegal or privacy-infringing material, for example. And the gravity of the risks of data sharing, both in terms of privacy and in terms of inadvertent amplification and "trophy" effects, may become less salient as time passes and public attention shifts.

But even with the distinct advantages of long-term archival approaches established, and its risks somewhat mitigated relative to short-term data sharing, the need for careful access controls—and the spirit of the Giftschrank—remains. Making archives like those contemplated above publicly available will likely never be palatable to the platforms—de-identification technology capable of preserving the usefulness of data is not foreseeable, and mirroring and aggregating content deemed unfit for inclusion on a platform stands to defeat part of the purpose of content moderation.

Theorizing the usefulness of archival professionalism and the Giftschrank to contemporary platforms is not the same thing as pointing towards an adoptable model. For that, we must dive into specifics and use cases.

II. DESIGNING A PLATFORM GIFTSCHRANK

PUT BRIEFLY, THE OBJECTIVE of a platform Giftschrank would be to preserve and control access to data about content moderation practices that couldn't otherwise be made accessible to present or future researchers. It would represent an option of last resort for carefully preserving sensitive material that would otherwise be lost.

The preceding section began to describe how a platform Giftschrank—and a long-term archival approach in general—might help accomplish this aim. We now explore some of the questions and considerations that platforms wanting to implement a Giftschrank model would need to address, particularly with regards to the scope of the Giftschrank's contents and controls on access to those contents (whether administered by the platforms themselves or by third-party archival management partners, like leading libraries). We will then explore two concrete use cases for the Giftschrank model—COVID-19 misinformation and Facebook's new Oversight Board.

A. Design considerations

The Giftschrank is much more a set of principles than it is an off-the-shelf, deployment-ready system. As discussed previously, the specific form and function of Giftschränke has varied significantly throughout their history, from locked cabinets to more subtle allocations of copyright and configurations of card catalogs. In adapting the model to their own purposes, platforms would need to make a distinct set of implementation decisions.

1. DATA COVERAGE

We have referred so far to a generalized form of Giftschrank for harmful content: an archive containing a comprehensive record, or highly representative sample, of all content subject to moderation actions along with metadata capturing the specifics of those actions. While, with accountability purposes in mind, such a general approach would offer a maximally robust paper trail around when and how a platform has used its moderation powers, the scale of online speech means that seeking to capture and preserve *everything* in a useful way may represent an overambitious initial undertaking. As such, platforms may consider deploying the Giftschrank model on more tightly

scoped subsets of the harmful content they address.

These subsets could be event-specific—capturing content and context surrounding particular crises, elections, or campaigns of abusive behavior, say—or they could be focused around particular policy changes or enforcement strategies. Upon rolling out a new content moderation standard, for instance, a platform could assemble a Giftschrank of all enforcement actions related to that standard, which could then be reviewed by internal and external researchers and auditors as a targeted accountability measure. The selection of these subsets would be a curatorial and investigative practice—one which might benefit from the hiring of internal archival teams that could support the scoping and indexing of materials for the benefit of researchers.

While such an approach could cover the full gamut of content handled by platforms—from hate speech to copyright violations—different access, retention, and collection rules may be required for different types of content. (And the archiving of some types of problematic content, like that which violates copyright, may only be feasible given special legal immunities.) Platforms would also have to determine what information and level of detail to include in any given Giftschrank. A platform might, for example, develop a sweeping archive of all of the enforcement actions it undertakes, but provide only a minimal collection of tags and metadata attributes for each record—like a more granular transparency report with circulation limited to approved researchers. This comprehensive accounting could be supplemented by much more detailed event-specific archives of the sort described above, which would provide greater resolution (perhaps including the content itself) on narrower areas of interest—like enforcement actions related to suspected election-year influence campaigns, or to a new policy banning attempts to unduly undermine faith in the integrity of balloting, or to an algorithmic tweak demoting health information flagged as misleading.

Of course, platforms should not be alone in determining what should be preserved as part of these archives. Rather, they might develop new research committees, comprising internal research and policy enforcement staff, as well as external researchers, to take on the curatorial role, developing strategies around which data to preserve. This process should be accompanied by frequent public readouts, welcoming comment from a wider range of researchers and members of the public. It could also be useful in working

through methodological questions, including whether a comprehensive capture of all content falling under the scope of a Giftschrank must be preserved, or if a representative sample would suffice.

Having determined the scope of examples and data attributes to be included in a Giftschrank, a platform would programmatically capture and store relevant records as part of the content moderation pipeline, assembling the resulting records within what archivists call a "dark archive"—a set of materials intended to be accessed only in the future, with any access constrained to its custodian until that time.[76] At this point, redactions and other curatorial actions could be applied, preferably by teams of archival professionals employed by the platforms, or by partner organizations capable of taking on that role. In some instances, as when the sensitivity of records within an archive is tied to one time-constrained event like an election, it may make sense to keep Giftschränke in this "write-only" mode for some period of time before making them available to anyone.

2. ACCESS CONTROLS

While determining the proper content of a Giftschrank would involve weighing complex policy considerations, the most challenging aspect of the design process concerns what happens once archives have been assembled and readied for use. Constrained access—the vetting of would-be researchers—is a cornerstone of the Giftschrank model, the feature that makes it possible for Giftschränke to hold highly sensitive materials.[77] Platforms would need to develop schemes for determining who should get access to a Giftschrank, as well as the terms of use associated with that access. To do so while advancing the purposes of the Giftschrank—accountability, transparency, and the furtherance of worthy research in the public interest—they would need to navigate a complex set of balances.

On one hand, platforms would be strongly incentivized to keep the number of researchers with access to a given Giftschrank relatively small. Each researcher granted access to a Giftschrank is another person who might infringe privacy rights while working with data that can't be fully de-identified, improperly leak harmful content back into the wild, or irresponsibly sensationalize data. While offering access to a larger number of researchers would provide better research coverage, narrowing access to a small handful

of reputable people and institutions would ensure a far greater degree of accountability and control. Participating researchers could be made subject to technical audits—depending on the specifics of the mode of access—vetted for trustworthiness and adherence to professional standards, and bound by the norms and expectations of their professional communities. And so long as participating researchers remain fully at liberty to publish their findings however they like, subject to the usual rigors of data use agreements and institutional review board approval, this trustworthiness could also be helpful in ensuring that those findings are responsibly narrativized in academic publications and the press. Participants could even review one another's methodologies and findings, perhaps as a condition of access, granting platforms an assurance of care without requiring them to go hands-on with independent research.

But this tendency towards restrictiveness could easily go too far, and, in any case, will raise thorny questions.[78] With small, highly empowered in-groups (likely drawn, in this case, largely from the academic elite), comes the potential for bias, failures in representation, and even cronyism.[79] If the group of researchers with access to a Giftschrank is not perceived as representative, neutral, and public-interest oriented, the usefulness of that group's research may be severely undermined. Indeed, questions relating to the membership composition of external bodies have plagued technology companies in the past.[80] Understanding and defining researchers as those who have affiliations at universities is also a U.S./EU-centric view of who qualifies professionally to conduct research studies. Balancing access and validating credentials as an access control continues to be an area ripe for additional thought and study.

To strike this balance, platforms should look to governance architectures that establish an appropriate degree of separation between access parameters and platform interests. They might conduct researcher selection processes in partnership with established research organizations and professional groups that specialize in developing and applying research standards. Or, to go a step further, they might even house Giftschränke with organizations other than themselves, transferring custody to established archival organizations. Leading libraries, for example, could store and provide access to Giftschrank data, managing curation and access as independent third

parties. Such a solution would provide maximal accountability at the cost of less direct control—consigning sensitive data to entities explicitly operating in the public interest—though platforms could surely condition their transfer of custody on continued involvement in archive management.[81]

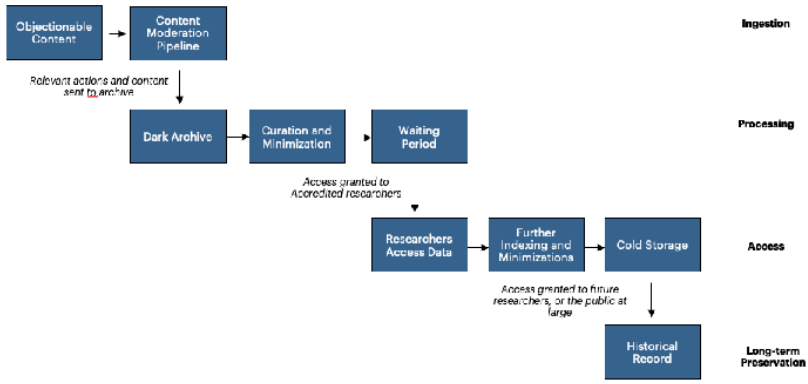
A third-party approach to the creation, maintenance, and administration of Giftschränke may also help address one of the key problems raised by a Giftschrank approachcost-based barriers to participation. The design and implementation of Giftschränke would involve both immediate and long-term costs—supporting development time, legal and policy support, researcher management, the creation of new infrastructure—which only a small handful of platforms would be able to absorb. Developing a common infrastructure for Giftschränke in the hands of, say, a consortium of leading libraries could open the door to more inclusive funding models, economies of scale, and the industry-wide formulation of best practices. Today's dominant platforms could still take the lead in funding the development of the Giftschrank model, but their work would form a basis for industry-wide progress.[82] Government could incentivize this accountability-maximizing third-party approach by offering protections or safe harbors to libraries and archival institutions offering their services.

Platforms and their partners would have to parameterize the conditions of researcher access, including specific restrictions on how data—particularly privacy-sensitive data—should be handled and used, mandated peer review arrangements, and limitations on the retention of data (by researchers, and by the platforms themselves). This would include defining the mode of access for researchers—platforms or third-party archivists storing Giftschrank data could, under a maximally restrictive approach, require accredited researchers to be on-site for data access. Alternatively, researchers could be required (where methodologically appropriate) to query Giftschrank data through a statistical analysis platform, such that raw data never changes hands. Queries could be audited regularly to prevent abuse.

And while it may make sense in some cases to set temporal retention limits on the contents of Giftschränke, platforms and their partners should also consider how Giftschrank material should be translated into contributions to the historical record, whether in part or as a whole. If and when a period of controlled research access expires, entire Giftschränke could be put into

"cold storage," preserved as dark archives for the benefit of future researchers.[83] If this long-term preservation proves overly expensive or infeasible, platforms might instead choose to store "samples" from a Giftschrank, or key data sets undergirding analyses produced and published by researchers.

In designing such systems, the architects of a Giftschrank should look to other areas of research involving the management and sharing of large volumes of sensitive data. For example, a massive literature (and regulatory apparatus) exists around guidelines and standards for the sharing of clinical data for medical research. This work, carried out over the course of decades, has strengthened professional norms in the medical research community and, despite continued debate on key issues, provides those engaged with such research with a strong foundation of best practices.[84] Also of note is the fact that the field of clinical research has made great strides towards establishing data sharing as a basic requirement of publication—cementing the responsible flow of sensitive data between researchers as a best practice in itself.[85]



Mapping one potential Giftschrank pipeline from collection to sharng

B. Giftschrank use cases

The question at the heart of the Giftschrank model is not just *how* a system should provide for greater transparency and research data availability, but

also *why* it should do so. Indeed, access to greater information about governance processes has sometimes proved a force for cynicism and polarization. Scholars like Lawrence Lessig[86] and David Pozen[87] have argued that transparency measures can hamper efforts to reform the very institutions they are designed to demystify and hold accountable. Building on the arguments for transparency in content governance already presented here, this section explores several concrete use cases for the Giftschrank concept, each tackling a particular area of platform policy in which accountability and data access are of paramount concern—and could be greatly buttressed by an archival approach.

1. COVID-19 MISINFORMATION

The current pandemic has ushered in a global wave of misinformation, from misleading medical advice and spurious accounts of impending government action to conspiracy narratives claiming, among other things, that the virus is a hoax, originated in a lab, or is caused by 5G towers.[88] Journalists, researchers, and public health agencies have sought to track and address this misinformation with varying degrees of success. Their efforts have been accompanied by expanded calls for data releases and transparency from platforms,[89] which are facing down their own set of challenges: the need to adapt and expand definitions of harmful content to address what could be imminent harms, staffing shortages and disruptions caused by the shift to remote work,[90] and knock-on effects that have had a deleterious effect on users' mental health.[91]

The pandemic has been a time of on-the-fly adaptation and triage for the platforms, which have adopted more aggressive content moderation strategies, leaned increasingly on automated moderation tools, and established new partnerships with public health organizations and other credible brokers.[92] Those public health organizations have had to adapt as well, leveraging social media to unprecedented extents to circulate credible information and debunk false narratives. It has also been a contentious time, during which the platforms have faced enormous scrutiny—Twitter, for instance, has found itself in the middle of a dispute around the propriety of its efforts to restrict misleading narratives posted by heads of state.[93]

COVID-19 misinformation represents a test of social media platforms'

ability to adapt speech controls to meet the distinctive demands of a crisis situation. And while many platforms have sought to be communicative in their handling of the crisis, a Giftschrank approach could establish a much clearer, richer, and more accountable basis for public understanding of platform responses. Beyond serving that core accountability interest, a Giftschrank preserving a reliable record of how platforms are moderating content—and what content they're moderating—could be useful to researchers, public health organizations, and the platforms themselves in planning for future crises of information quality and access.

Timescale considerations also agitate for a Giftschrank for COVID-19 misinformation. The pandemic represents a world-historical event—a moment in history that scholars, policymakers, and members of the public will look back to time and time again. It stands to be a touchpoint for any future health care crisis, and particularly any global-scale crisis where the efficacy of a response is largely dependent on the circulation of credible information. The usefulness of data about COVID-19 misinformation therefore likely has an exceptionally long shelf-life, and serves to fill what would otherwise be a serious gap in future understandings of the pandemic. That long shelf-life suggests a long-term archiving approach.

And a lengthy timescale would actually make it possible for platforms to take immediate action. Platforms could start constructing a write-only "dark archive" precursor to a Giftschrank immediately, with the expectation that more complicated and deliberative archive preparation processes—like minimization, indexing, and other measures—would be considered and carried out once the heat of the crisis has passed. In that sense, the Giftschrank approach would actually buy the platforms time, enabling them and their archival experts to make a specific decision on the form and function of an eventual archive—or even whether to go through with the archive at all—in a less chaotic moment.

Indeed, that option to securely capture first and build archives later means that the Giftschrank model is perfectly suited for moments of crisis, avoiding many of the risks, time pressures, and resource constraints that a real-time data sharing approach would impose. The COVID-19 pandemic may also provide an opportunity for a multi-platform Giftschrank, given that misinformation has manifested across different platforms in different

ways. A single collective archive, perhaps managed by a trusted third party like a library consortium, could provide future researchers with the material to draw significant comparative and cross-cutting insights.

2. THE FACEBOOK OVERSIGHT BOARD

Facebook's newly minted Oversight Board for content decisions has the potential to cement itself as a bold experiment in accountable content governance. The board will have the opportunity to review takedowns that have been appealed by Facebook users—or forwarded to the board by Facebook itself—in a binding way.[94] The board's decisions on appealed takedowns will both determine the outcome of the case under consideration and set precedent, such that they will shape future content decisions of a similar nature. If it can meaningfully bind corporate action to the judgments of experts beyond the walls of Facebook itself, the board stands to model a new way of thinking about platform accountability around content moderation decisions.

As an institution intended to build legitimacy and public buy-in around Facebook's content policy processes, the Oversight Board's success is largely dependent on its credibility and transparency. For the board to be perceived as a legitimate governance institution, Facebook users and members of the public will need to be convinced that it functions autonomously, makes coherent decisions, and has the power to meaningfully affect the course of Facebook's policy. To support this long-term legitimacy-building process, Facebook might consider implementing two related Giftschränke—one covering the content and other "exhibits" coming before the board, the other tracking the implementation of specific board decisions across Facebook.

The first Giftschrank will be essential in establishing—now and into the future—the integrity of the board's review process. Researchers and others seeking to understand the course of the development of the board's "jurisprudence"—and its operational conventions, which will surely shift substantially as it grows into its role—would benefit from access to the materials that the board has before it in making determinations on cases. While some of these materials may be sensitive enough that they cannot be shared publicly, preserving and archiving them in an organized way will ensure that the board's decisions can be fully contextualized well into the future. The mere fact of such an archive's existence may be credibility-building in

itself, ensuring that reference is possible under exceptional circumstances.

Such an archive is particularly essential given that Facebook is looking to the board as a source of precedent. Drawing on precedent often means understanding why past decisions were made, considering their scope and specificities, and even assessing whether they might have been wrongly rendered. In its most limited form, the Giftschrank could be accessible only to members of the Oversight Board itself, such that they could review the full context of past decisions. But by formalizing the archive even for this limited audience, Facebook would be developing a coherent record of board actions that could eventually be opened up—in part or in full—to future researchers, or even to the public at large.

The usefulness of the second style of Giftschrank would pick up where the first leaves off. Decisions made by the Oversight Board are supposed to translate into policy adjustments by Facebook, but the means of this translation are not entirely clear. To ensure accountability, the board might be given the opportunity to charter new Giftschrank archives meant to capture Facebook's enforcement of its rules around particular types of content decisions. If, for example, the board made a ruling requiring Facebook to leave up groups for organizing anti-quarantine protests, it might request that an archive of all actions directed at anti-quarantine pages be created. This ingestion of materials may be more feasible in some cases than in others, where takedowns relevant to a decision are less separable from irrelevant takedowns.

In cases where they are feasible to implement, archives of this second kind would allow the tracking of board decisions between theory and practice, ensuring accountability around Facebook's implementation of the board's will. The archives might be audited by independent subject-matter experts appointed by the board itself to follow up on high-stakes decisions that the board feels should result in significant changes. In instances where mismatches materialize, this review process would provide an opportunity for further dialogue and deliberation between Facebook and the board. In instances where board decisions and Facebook's practices align, the board may be able to report its empirically substantiated successes to the public, bolstering confidence in its ability to meaningfully shape policy.

Both of the Giftschränke would also serve a longer-term purpose—providing

the material for a more comprehensive accounting of the board's development, including its stumbles. Such an accounting, particularly if produced by independent researchers, could provide a credible basis for the development of future oversight boards with similarly binding power, helping to convert a single-platform experiment into an industry-wide best practice.

III. BARRIERS TO THE GIFTSCHRANK MODEL

Even when all stakeholders agree conceptually that a data-sharing approach may hold potential as a remedy to a problem of urgent concern, implementation within the context of a private platform or firm more generally encounters barriers, including a potential first-mover disadvantage, legal exposure, and privacy risk. This section examines practical considerations that might forestall adoption of a Giftschrank model.

A. The first-mover disadvantage

There is often a first-mover disadvantage among private sector peers considering new transparency mechanisms.[95] While there may be a general understanding or assumption that activities like hate speech occur across multiple platforms, hard data still has the power to shape public narratives. A first-mover may well end up furnishing concrete proof of objectionable activities on its platform in particular—and when researchers only have one data source, contextualizing platform-level problems in terms of ecosystem-level trends means speculating. What's more, companies that invest in transparency mechanisms despite the risks rarely receive much credit for these efforts, especially if the phenomena exposed are ugly or shocking. It's understandably difficult to praise a company for putting a Pandora's box of its worst elements, including previously unseen ones, on public display, especially if others keep their problems better-hidden. Further, if one company tries a transparency model that results in public relations blowback, other companies can avoid investing money, time, and risk in such an approach, even if industry-wide adoption would result in better optics and outcomes.

However, it can also be argued there is a competitive advantage to

making public commitments, and being a corporate leader with external stakeholders (e.g., civil society advocates and academics) in creating accountability mechanisms that serve larger social interests. The first mover can define the terms of retention and access, and do so in a way that limits other liability risks. The time delay component is also a vital element of the Giftschrank model, in that it can provide a degree of separation between platform decision makers and future scholarship. For certain very sensitive material, archives could be kept "dark" for decades, released long after the sensitivities have attenuated, and, for internal incentives purposes, after most current employees have retired and restricted stock units are vested. Of course, this same time delay is likely to also instill public distrust in ultimately upholding these commitments, which is why other design elements and governance are included so that private firms could not issue false promises.

B. Privacy risks and legal vulnerabilities

Data sharing and information access creates regulatory uncertainty for private companies. Assessing privacy risks has been a focus of scholarship throughout the last several years in order to better encourage private companies to engage in responsible data sharing.[96] A model like the Giftschrank opens up unknown liability for a company, especially as laws stand to evolve over the lifetime of a Giftschrank and its materials. It is possible that relevant statutes of limitations could expire for causes of actions based on material set aside for a Giftschrank, therefore limiting the potential liability of a private company.

Within the landscape of global laws, the intricacies of managing the legal threat exposed by the Giftschrank model remains overwhelming. From a legal risk perspective, creating any public (or semi-public) archive, complex or extensive data sharing or access mechanism (data sharing here is viewed as implicit with the contents of the archive), or documentation of politically and ethically fraught material may be irrational. The risk of regulatory change, liability, or bad press stands to outweigh possible benefit to a private firm, especially when such risks implicate unknowns and cannot be easily quantified. For instance, one cannot anticipate all of the means by

which someone could re-identify or otherwise harm individuals whose data is included in an archive. It is also possible that de-identification methods and privacy-preserving techniques used to help protect privacy interests at the creation of a Giftschrank would be made obsolete or less protective by the time that data is made accessible, placing further weight on the trustworthiness of the researchers chosen to receive access.

One existential design threat is that the Giftschrank could be abused to capture speech deemed objectionable by an authoritarian government and reverse engineered to harm individuals. Or, speech could be captured and later re-identified in a way that documents past beliefs and harms a person in the future. These privacy harms underscore why the design of privacy preserving measures are vital to enabling a Giftschrank model.

Further, no one employee (or handful) would want to be responsible for signing off on so many unknowns, and without industry-level archival standards or safe harbors it would require significant outside support and internal leadership buy-in for a company to commit to launching a Giftschrank model. That said, decisions to share data despite extensive risk and expense are not without precedent: the Social Science One Initiative at Facebook attempted to create at-scale data sharing that implemented differential privacy techniques to enable unprecedented research to occur on misinformation around elections. This initiative required significant investment and commitment from a private company (Facebook) alongside external partners to implement. Though it has experienced both criticism and setbacks, it illustrates that—with sufficient corporate buy-in, support, and investment—ambitious public-interest data sharing initiatives can make it off the launch pad.

A Giftschrank would need to be designed to be consistent with obligations made to users about how their information could be used or shared at the time of their post. Preservation of data generated by users will also invoke concern around compliance with international privacy laws. The propagation of data protection interests and vindication of legal rights after data has been shared with a third party, regardless of purpose, is one of the key challenges of the current data protection era. There are specific concerns around complying with data use and retention obligations within the EU's General Data Protection Regulation (GDPR)[97] and other new and evolving

legal regimes, like the California Consumer Privacy Act.[98] Retaining data for such a long time is likely to frustrate current norms around user rights to delete past posts or even specific legal obligations like GDPR's right to erasure (commonly known as the "right to be forgotten"). The time component of the Giftschrank adds further complications: if a law is passed that allows users—even those who propagate hate speech—to request the deletion of their data, or to review data that is kept about them, it could complicate retention in the Giftschrank. A user-driven redaction model would, at best, require sophisticated and ongoing alterations to Giftschrank contents.[99]

Depending upon the design of the Giftschrank, it is possible that posts may be de-identified so that some user privacy concerns may be minimized.[100] This could impact the quality of the dataset or ability to look for patterns among posters or social networks. An upside to the time variable in the Giftschrank is that it would be harder to re-identify posts by searching the contents on a search engine or platform—a known risk in social media research when tweets or posts have been sanitized by removing the user handle.[101] By waiting many years, it is likely these posts could be taken down (possibly by content moderation itself—though perhaps not uniformly across sites) or simply become obfuscated through time.

However, some posts may be discussed in blogs, research papers, and the news, which would make it difficult to protect the identity of the poster in the future. Many could argue that someone who posts hate speech does not have a right to privacy, or that by posting the speech in a public fashion vitiated their right to privacy. Yet, internet policy is often defined by fringe cases. There are ethical balances to consider. For instance, if an 18-year-old posts hate speech, and later regrets and takes down the text (see above for discussion on considerations around the right to be forgotten), this could be forever enshrined within a Giftschrank archival model. These ethical concerns may be minimized and considered within the implementation design, and by limiting identifying information, researcher access, and codes of conduct for researchers using these data in the future.

Giftschränke do not necessitate global coverage and access. Given the growing complexity of international laws on content, privacy, and misinformation, a Giftschrank (or network of Giftschränke) should be kept regional to lessen legal liability, complexity, and exposure. This would likely privilege

particular regions over others, and give them access to future research likely to have contemporary salience. But there could be other ways of expressing priority (e.g., political or real-world necessity) in determining which regions are included in an implementation model.

Another possible legal exposure would be that if data are retained and held for future access, governments or other actors could possibly sue companies to gain access to these archives before they were designed to be released. This would not only frustrate privacy and ethical concerns designed to be addressed by locking content away for a set time period, but also would create another risk and disincentivize adoption. It is possible this risk could be negated if a third party held the archive, or if there were safe-harbor-style laws passed in a region to protect information stored in an archive like a Giftschrank.[102]

Where there are privacy risks, there are often security risks. A digital implementation of a Giftschrank would have a large attack surface (both in terms of data included and length of time stored) that could be attractive to bad actors. There is a unique threat model associated with a digital Giftschrank that collects sensitive material in one location for long-term storage and centralized access. This creates security risks and further liabilities for companies.

The design of the Giftschrank could play a role in minimizing legal exposure and risk, and adding in additional protections for privacy and ethics. Limiting access (e.g., to only accredited researchers), putting limitations on use and access through the use of clear rooms or other gating and control mechanisms, and thinking about who should govern the Giftschrank itself should be considered to minimize risk and legal vulnerability.

IV. STARTING POINTS FOR GIFTSCHRANK IMPLEMENTATION

Previous sections outline an idealized model for Giftschrank-style archival implementation in the platform era. Given possible misalignment in incentives and privacy risks as barriers, an all-or-nothing approach to implementing a Giftschrank model from scratch

would be unrealistic. Here we propose partial implementations and pragmatic places to start a Giftschrank that attempt to minimize barriers for private firms.

A. User-driven data donation models

The most lightweight—and least platform-dependent—approach to creating a Giftschrank for online content could be to crowdsource collection through dispersed data donation models that avoid corporate incentives and action altogether. Even if private companies want to commit resources to a Giftschrank-style initiative, it could still take years or be held up (or altered) at various points, depending on external pressures and legal uncertainty. Corporate participation and external, crowd-driven initiatives do not have to be mutually exclusive either: It may be possible to design the two types of Giftschrank such that each complements the design and limitations of the other for a more robust accountability system overall.

Users and researchers already interact—to varying degrees—with hate speech and other harmful content online through the nature of their digital lives and activities. Though scraping and systematic collection of content is often prohibited by terms of service on public platforms (in part due to privacy concerns around Cambridge Analytica-style mass harvesting of user data), it is possible that smaller-scale curation could be done in a way that collects key examples without mass violation of privacy rights or legal terms. The post-WWII German Giftschrank of Nazi texts was not an exhaustive collection of every instance of hate speech ever recorded, but rather a curated collection meant to encapsulate enough of the rhetoric needed to enable future study and analysis. Using even small-scale, selected captures to document the lived experience of users could enable a curated collection of particular kinds of hate speech. Platforms could explicitly support donation, including natively within their abuse reporting tools, and set standards for who can do so—either only selected researchers or perhaps a larger crowdsource. Here, too, librarians and archivists should have a role—both in curating content within the archive, and in instilling some agreed-upon level of data cleaning and possibly other privacy-preserving measures.

These user donation models for small-scale data collection may risk decreased data quality and less robust access controls than a more

centralized and "official" Giftschrank model. And there is reason to believe that the legal and ethical concerns attending a platform-led data access model may persist—indeed, the field of medical research is currently engaged in a fierce debate over the propriety of soliciting, using, and sharing data explicitly donated by patients.[103] These considerations may be mitigable by integrating members of the library and archival professions at the site of the crowdsourced infrastructure—another reason to engage with specialist third-party institutions like libraries early and often, regardless of the ultimate form a Giftschrank takes.

The Giftschrank model could also be amended to preserve researchers' individual data sets used for the small-scale study of misinformation and hate speech. Researchers who have already hand-curated samples of hate speech could have a platform to donate their collections into long-term preservation so that future researchers could better benefit from their work. There are many possible models that could work with (but not depend upon) private companies to be the initiators and designers of Giftschränke. Such models could leverage cooperation and support from private companies by enabling them to participate in external design and implementation, even without official action. An externalized model focused on cooperation with platforms could bootstrap long-term data sharing by avoiding a collision of interests and conflicts around how public data is captured, stored, and shared.

Platforms could also enable larger-scale, user-directed sharing through rights like data portability. Data portability enables users to transfer their personal information from one digital context to another on their own prerogative. Currently, these data portability tools are in their infancy and limited to instances where it is clear whose data is being ported, and where privacy risks have been minimized or eliminated. There are key questions, however, about the data protection interests of other users when ported information involves multiple parties—as would be the case for hate speech comments or possible public posts within a social network. Though current portability tools are not now designed to port content that might include hate speech (unless of course that speech was made by the directing individual), it is possible future tools could include comments and other data types that would be useful for a Giftschrank.

B. Cooperative archival governance

In order to better address concerns around first-mover disadvantage—or to possibly instill peer pressure between companies—a cooperative commitment could be formed with a public set of companies to contribute or help facilitate the development of industry-wide archiving practices. This model would require external leadership and governance, but member companies could possibly contribute funding, as well as internal tools or expertise to feed into a cooperative archival approach. Such a model could be used to generate shared responsibility between a third-party data holder and data providers, and enable a Giftschrank in part through the equal distribution of risk.

Collaborative governance would fall on a natural spectrum of involvement. At one extreme, platforms could consolidate their technical infrastructures, unifying the Giftschrank "stack" across private-sector actors. At the other, they could simply develop industry-wide best practices, messaging, and research access protocols.

Though it does not implicate the sorts of access control measures envisioned for a Giftschrank, the Lumen project at Harvard University's Berkman Klein Center for Internet & Society provides one model for industry-wide data sharing around content takedowns. Every day, Lumen collects and indexes thousands of "requests to remove materials from the web"—many of them submitted under the notice-and-takedown process of the Digital Millennium Copyright Act (DMCA). These requests are shared with Lumen by a broad range of contributors including Google, Twitter, YouTube, Wikipedia, and Stack Exchange.[104] The more than 10 million notices included in Lumen's database have formed the basis for significant research projects, including a research project from UCLA Professor Eugene Volokh on fraudulent takedown notices[105], and a Wall Street Journal analysis of DMCA-driven link suppression by Google users. (After the Journal shared its results with Google, the platform reportedly restored upwards of 52,000 improperly removed links.)[106]

By putting in place consolidated collection and access infrastructure for takedown data, Lumen has developed a platform for ecosystem-level analysis of takedown request patterns. New contributors are liberated from the burden of envisioning a bespoke sharing protocol for data relating to

takedown notices—they can instead simply plug into the project's tried-and-true approach.

Regardless of the specific form of any collaborative archival structure, external coordination and oversight will be essential. As mentioned previously, a consortium of libraries or other civil society partners may be in a position to provide a coordinating platform for such an undertaking, whether by administering data infrastructure or by providing guidance, oversight, and archival expertise. Such a model would make it far easier to onboard new private companies wanting to contribute to a Giftschrank, both by formalizing Giftschrank protocols and by providing an authoritative statement of best practices. It would also provide a basis for greater public accountability, pulling certain Giftschrank functions out from behind the corporate veil.

C. Regulatory measures to enable (or require) Giftschränke

As mentioned in the previous section, legislation could help boost incentives and alleviate risks associated with private firms' creation of or participation in a Giftschrank model. For instance, a law creating a safe harbor for the content donated or stored in an archive would significantly mitigate legal risks inherent in Giftschrank adoption. Given that platforms' willingness to retain and share data is significantly restricted by such barriers—or the possibility that such risks might later materialize—a legislative assurance of legal protection for Giftschrank-style archiving, subject to strict conditions, might enable otherwise infeasible forms of corporate action.

Given the significance of the accountability interests served by archiving, legislators may even go a step further, setting out transparency and research access requirements by which private platforms of a certain size should abide. These requirements could include, for example, specified retention periods for certain data related to content removed by platforms, a mandate to sign onto or develop a data-sharing framework within certain parameters, and the regular publication of information regarding how data is being shared and used for research purposes. In most cases, these requirements would need to be accompanied by safe harbors and regulatory aids of the sort described above. Such an approach, though fraught with far too many questions and potential risks to be explored in depth here, could get around

many of the disincentives to data sharing confronted by today's platforms, particularly if designed with enough flexibility to allow platforms to tailor implementation to their own needs.

This carrot-and-stick approach could be formulated to steer platforms towards a data sharing and transparency model built around third-party archival partners drawn from civil society. As discussed above, such an approach would facilitate the development of best practices; provide a degree of independent accountability around the design and implementation of the Giftschrank model; and bring to bear the native expertise of the sorts of institutions, like libraries and archival organizations, that would best fit the role. Distributing responsibility for archiving beyond the walls of any one firm would mean establishing the Giftschrank not just as another platform initiative, but as a real partnership between industry, civil society, and—through its enabling regulatory powers—government. This sort of multi-sectoral arrangement, if given real legal weight, might provide even those deeply critical of platforms' data sharing efforts to this point a reason to keep an open mind.

Outside of questions of legal immunity, public regulators could further increase the viability of a Giftschrank model by clarifying—and limiting—how social media companies' archives could be accessed under warrant by law enforcement and intelligence agencies. Given the civil liberties risks implicated by such access, firm rules of the road would be essential, along with measures to ensure the consistent tracking and reporting of government access requests.

Many regulatory approaches to driving Giftschrank adoption may risk significant conflicts with other values discussed within this essay, as well as incompatibility with privacy policies, security mandates, or responsible data management practices. At its worst, regulation could create impossible mandates that collide with other areas of law and policy. Legislative proposals also may narrowly prescribe regional approaches that conflict with global standards, or could be drafted by governments with the intent of creating repositories that invade citizen privacy and impede free speech. This paper elucidates the need for careful policymaking that incentivizes public interest action, but prevents overly prescribed and under-designed Band-Aid approaches that risk privacy and security harms.

V. CONCLUSION

THE ARCHIVAL APPROACH described here is by no means a catch-all solution to the problems faced by platforms seeking a more transparent and accountable record of moderation actions. Giftschränke cannot and will not replace efforts to develop effective protocols for real-time data sharing, nor do they represent truly "open" resources, centered as they are around access control schemes. But even so, they may be our best alternative—at least for now—to a paradigm in which it makes little sense for platforms to do anything other than to delete first and ask questions later.

Prospects of implementing a Giftschrank face real challenges, many of which we have laid out here. But part of the ethos of archiving is that every little bit counts—whether by considering some of the initial schema laid out here, by constraining a pilot Giftschrank to a very narrow topic, or by attempting a dry-run dark archive without making any representations to outside stakeholders. In this way, platforms and their archival partners can start today in exploring what a better-documented future for content moderation might look like. The Giftschrank model raises important questions and challenges that should be further studied and developed by practitioners and researchers.

NOTES

1 Caesar C. Aronsfeld, *Mein Kampf*, 1945-1982, 45 Jewish Soc. Stud. 3, 311 (1983).

2 Stephen Luckert, *Mein Kampf Enters the Public Domain*, Atlantic Monthly, (Dec. 31, 2015), https://www.theatlantic.com/international/archive/2015/12/mein-kampf-copyright-expiration/422364/ [https://perma.cc/3DYZ-NRZR].

3 Sam Greenspan et al., *The Giftschrank*, 99% Invisible (Mar. 8, 2016), https://99percentinvisible.org/episode/the-giftschrank/transcript/ [https://perma.cc/5ENR-6VK4].

4 For a more robust English-language treatment of the history of the Giftschrank in the wake of World War II, see Jennifer Allison, *On Censorship: Lessons from the Giftschrank* (2020) (unpublished manuscript) https://works.bepress.com/jennifer_allison/81/ [https://perma.cc/4MSR-BPRY].

5 Der >Giftschrank<: Erotik, Sexualwissenschaft, Politik und Literatur: >Remota< Die weggesperrten Bücher der Bayerischen Staatsbibliothek 9-21 (Stephan Kellner ed., Bayerische StaatsBibliothek 2002).

6 Greenspan, *supra* note 3.

7 Aronsfeld, *supra* note 1, at 311-315.

8 Peter R. Range, *Should Germans Read 'Mein Kampf'?*, N.Y. Times, (Jul. 7, 2014), https://www.nytimes.com/2014/07/08/opinion/should-germans-read-mein-kampf.html [https://perma.cc/BH42-EJC7].

9 To be sure, the restriction or censure of broad categories of speech is far less familiar to the American legal tradition than to the German (let alone that of 16th century Bavaria, where the Giftschrank originated). But today's online platforms, the principal subjects of this paper, can restrict speech far more extensively than can public authorities in either country, bringing the sort of purging considered here within the realm of feasibility.

10 Randall C. Jimerson, *Archives for All: Professional Responsibility and Social Justice*, 70 Am. Archivist 252, 252-281 (2007).

11 Alison Flood, *New Edition of Mein Kampf Set to Land on German Bestseller Lists*, Guardian (Jan. 13, 2016), https://www.theguardian.com/books/2016/jan/13/mein-kampf-german-bestseller-lists-new-edition-adolf-hitler [https://perma.cc/HUZ3-UKVR].

12 Jonathan L. Zittrain, *Three Eras of Digital Governance* 1-9 (2019) (unpublished working paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458435 [https://perma.cc/WHP3-AN6N].

13 *See, e.g.*, Tarleton Gillespie, Custodians of the Internet (Yale Univ. Press 2018).

14 See, for example, Facebook's Transparency Reports, which give figures for the number of enforcement actions on different types of content on a quarterly basis. To give some illustrative examples, Facebook enforced its hate speech policies by taking action against 1.6 million pieces of content in Q4 of 2017. For Q2 of 2020, the figure stood at 22.5 million pieces of content. While patterns vary across enforcement categories, the number of enforcement events captured across the report has generally trended upwards. Facebook, *Community Standards Enforcement Report*, Facebook (2020), https://transparency.facebook.com/community-standards-enforcement [https://perma.cc/AX26-9ETJ].

15 *See, e.g.*, *Understanding the Community Standards Enforcement Report*, Facebook Transparency (2020), https://transparency.facebook.com/community-standards-enforcement/guide [https://perma.cc/J74V-LW7A]. (See section 4 on appealed and restored content.)

16 *See, e.g.*, *Information for Law Enforcement Authorities*, Facebook (2020), https://www.facebook.com/safety/groups/law/guidelines/ [https://perma.cc/E346-552L].

17 *See, e.g.*, Alex Warofka, *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*, Facebook (Nov. 5, 2018), https://about.fb.com/news/2018/11/myanmar-hria/ [https://perma.cc/4BSN-MF2N].

18 Facebook's Social Science One project offers one example of a "one-off" project. It centered on metering access to a single dataset, rather than building a pipeline for the development and sharing of archives with a community of researchers.

19 *See, e.g.*, Brandon Silverman, *CrowdTangle for Academics and Researchers*, Facebook for Media

Blog (Jan. 28, 2019), https://www.facebook.com/facebookmedia/blog/crowdtangle-for-academics-and-researchers [https://perma.cc/L4X7-BH3E].

20 *See, e.g.*, Vijaya Gadde & Yoel Roth, *Enabling further research of information operations on Twitter*, Twitter Blog (Oct. 17, 2018), https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html [https://perma.cc/7SBV-N4HG].

21 *See, e.g.*, Bethan John & Clea Skopeliti, *Coronavirus: How are the social media platforms responding to the 'infodemic'?*, First Draft News (Mar. 19, 2020), https://firstdraftnews.org/latest/how-social-media-platforms-are-responding-to-the-coronavirus-infodemic/ [https://perma.cc/BR3G-8P7B].

22 Society of American Archivists Core Values Statement and Code of Ethics, Society of American Archivists (Aug. 6, 2020), https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics [https://perma.cc/693X-DDVH].

23 John D. Bowers and Jonathan L. Zittrain, *Answering impossible questions: Content Governance in an Age of Disinformation*, 1 Harv. Kennedy Misinformation Rev. 1-8 (2020).

24 Der >Giftschrank<, *supra* note 5, at 6-7.

25 Der >Giftschrank<, *supra* note 5, at 9-21.

26 John D. Bowers & Jonathan L. Zittrain, *supra* note 23, at 1-8. *See also* evelyn douek, *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability*, 121 Colum. L. Rev. (forthcoming 2021).

27 Greenspan, *supra* note 3.

28 J.K. ROWLING, HARRY POTTER AND THE PHILOSOPHER'S STONE Chap. 13 (Bloomsbury 1997).

29 Historische Kataloge der Bayerischen Staatsbibliothek München 524 (Stephan Kellner & Annemarie Spethman eds., Harrassowitz Verlag 1996).

30 Greenspan, *supra* note 3.

31 Hugo Hayn, Bibliotheca erotica et curiosa monacensis: Verzeichniss französischer, italienischer, spanischer, englischer, holländischer und neulateinischer Erotica und Curiosa, von welchen keine deutschen Übersetzungen bekannt sind (M. Harrwitz 1889).

32 Der >Giftschrank<, *supra* note 5, at 18.

33 The extent to which the separateness of Giftschrank collections was actually enforced seems to have varied, admitting of some amount of play in the joints. *See* Jennifer Allison, *supra* note 4, at 8-9.

34 Greenspan, *supra* note 3.

35 Greenspan, *supra* note 3; U.S. Dept. of the Air Force, Air Force Instruction 91-104, Nuclear Surety Tamper Control and Detection Programs (2013).

36 Christoph Caspar, *'Mein Kampf': A Best Seller*, 20 Jewish Soc. Stud. 1, 3-16 (1958).

37 Greenspan, *supra* note 3.

38 Thomas Bürger, *Aus Dem „Giftschrank" In Das Internet? Ist Aufklärung Über Ns-propaganda Im Offenen Wissenschaftsnetz Möglich? Eine Tagung In Wien Zur Verantwortung Von Bibliotheken Und Museen Sucht Nach Neuen Wegen*, 73 Mitteilungen der VÖB 153, 153-157 (2020)

39 And indeed, a Giftschrank approach may raise particular censorship concerns in the digital context. With the centralization of information–particularly when that information is locked within a closed system like a Giftschrank–comes the opportunity for ad hoc, less-than-transparent modification and censorship. *See* Jonathan L. Zittrain, *The Internet's Fort Knox Problem*, TAP: Tech., Acad.'s, Pol'y (2010), https://www.techpolicy.com/Blog/June-2010/The-Internet%E2%80%99s-Fort-Knox-Problem.aspx [https://perma.cc/58ZD-R8J5].

40 Greenspan, *supra* note 3.

41 Catherine Plum, Antifascism After Hitler: East German Youth and Socialist Memory 1-2 (Taylor & Francis 2015). Much of this obscurity remained even as Germany reunited and the GDR's Giftschränke opened–after decades of operating in the Soviet model, libraries in the GDR lacked the infrastructure to effectively provide ready access to newly available resources. *See, e.g.*, B. Venkat Mani, Recoding World Literature 179-213 (Fordham University Press 2016).

42 John Milton, Areopagitica (Sir Richard C. Jebb ed., Cambridge Univ. Press 1918) (1644).

43 *Id.* at 7.

44 Much righteous censorship, but not all. Milton was, for all of his brilliance, a man of his time: "I mean not tolerated popery, and open superstition,

which as it extirpates all religions and civil supremacies, so itself should be extirpate." *Id.* at. 60.

45 *Id.* at 20.

46 *New York Times Co. v. Sullivan*, 376 U.S. 254, 279 n.19 (1964).

47 *See, e.g.*, *Facebook Community Standards*, Facebook (2020), https://www.facebook.com/communitystandards/ [https://perma.cc/JF8M-P8XY]; *The Twitter Rules*, Twitter (2020), https://help.twitter.com/en/rules-and-policies/twitter-rules [https://perma.cc/TFB4-YW9]; *Reddit Content Policy*, Reddit (2020), https://www.redditinc.com/policies/content-policy [https://perma.cc/Z2QE-UGJX]; *Content Policies*, Google (2020), https://support.google.com/news/publisher-center/answer/6204050?hl=en [https://perma.cc/9YSP-TKU7].

48 Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598, 1630-1657 (2018).

49 Travis Yeh, *Addressing Sensational Health Claims*, Facebook (2019), https://about.fb.com/news/2019/07/addressing-sensational-health-claims/ [https://perma.cc/6E2D-XEH5].

50 Elizabeth Dwoskin, *How Twitter Decided to label Trump's Tweets*, Wash. Post (May 29, 2020), https://www.washingtonpost.com/technology/2020/05/29/inside-twitter-trump-label/ [https://perma.cc/K6UG-Z53H].

51 Dieter Bon, *One of Twitter's new anti-abuse measures is the oldest trick in the forum moderation book*, Verge (Feb. 16, 2017), https://www.theverge.com/2017/2/16/14635030/twitter-shadow-ban-moderation [https://perma.cc/WT4N-HKT9].

52 Klonick, *supra* note 48, at 1658-1662.

53 *Facebook Community Standards §16: Cruel and Insensitive*, Facebook (2020), https://www.facebook.com/communitystandards/cruel_insensitive [https://perma.cc/28FH-BN4C].

54 *Facebook Community Standards §21: False News*, Facebook (2020), https://www.facebook.com/communitystandards/false_news [https://perma.cc/6ZY9-F3JF].

55 *Facebook Community Standards §12: Hate Speech*, Facebook (2020), https://www.facebook.com/communitystandards/hate_speech [https://perma.cc/T7LT-GC8W].

56 Mark Zuckerberg, *Facebook's commitment to the Oversight Board*, Facebook (2019), https://about.fb.com/wp-content/uploads/2019/09/letter-from-mark-zuckerberg-on-oversight-board-charter.pdf [https://perma.cc/2D2R-Q9WS].

57 Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 Yale L. J. 2418 (2020)

58 Klonick, *supra* note 48, at 1665-1666.

59 47 U.S.C. § 230

60 2018 Edelman Trust Barometer 19 (2018), https://www.edelman.com/sites/g/files/aatuss191/files/2018-10/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf [https://perma.cc/TJ4H-8PHG].

61 Danielle K. Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 Notre Dame L. Rev. 1035, 1035-1071 (2018). *See also* California Consumer Privacy Act Cal. Civ. Code §1798.100.

62 *See, e.g.*, Klonick, *supra* note 48, at 1654. (Describing an incident in which Facebook publicly committed to rewriting some of its rules around nudity and newsworthiness after it repeatedly removed an iconic Vietnam War photo showing a naked Vietnamese child fleeing a napalm attack.)

63 The COVID-19 pandemic has created a unique set of difficulties for platforms seeking to enforce complex rules. At the outset of the pandemic, platforms saw significant shortfalls in human review capacity and announced increased reliance on mistake-prone automated detection capabilities. *See, e.g.*, Matt Derella and Vijaya Gadde, *An update on our continuing strategy during COVID-19*, Twitter (April 1, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html [https://perma.cc/DY2G-PLC4].

64 Nick Clegg, *Welcoming the Oversight Board*, Facebook (May 6, 2020), https://about.fb.com/news/2020/05/welcoming-the-oversight-board/ [https://perma.cc/7H57-WS7F].

65 John D. Bowers and Jonathan L. Zittrain, *supra* note 23, at 1-8.

66 By "public faith," we mean not just the buy-in

of platforms' users, but also that of research communities, regulators, and non-users who encounter the platform through public debate and discussion. Given the stature that online platforms have assumed in contemporary social and political conversations, a narrower audience would necessarily miss key voices and decision-making constituencies.

67 Emma Llansó, *COVID-19 Content Moderation Research Letter–in English, Spanish, & Arabic*, Ctr. for Democracy & Tech. (Apr. 22, 2020), https://cdt.org/insights/covid-19-content-moderation-research-letter/ [https://perma.cc/N8G5-NPHP].

68 *GDPR and the Impact on Data Archiving and Information Governance*, Actiance 3-11 (2017), https://www.actiance.com/wp-content/uploads/2017/03/WP-GDPR-Impact-on-Data-Archiving-and-Information-Governance.pdf [https://perma.cc/7RJE-WZJG].

69 Camille François, *Actors, Behaviors, Content: A Disinformation ABC*, Transatlantic Working Group (2019), https://science.house.gov/download/francois-addendum [https://perma.cc/5ZHG-677E].

70 Interview by Oumou Ly with Claire Wardle, founder and director, First Draft, remote (June 4, 2020), https://medium.com/berkman-klein-center/the-breakdown-claire-wardle-on-disinformation-and-todays-journalistic-conventions-51ce5a41fe7d [https://perma.cc/7J52-2T2S].

71 *See, e.g.*, Neil Miller, Banned in Boston: The Watch and Ward Society's Crusade against Books, Burlesque, and the Social Evil (Beacon Press 2010). (The Watch and Ward Society, a New England literary censor, zealously banned media containing salacious content. Being "banned in Boston" became a mark of honor actively pursued by publishers, who would advertise their outputs accordingly.)

72 Jimerson, *supra* note 10, at 252-281.

73 SAA Core Values Statement and Code of Ethics, *supra* note 22.

74 *Id.* ("[Archivists] should embrace principles that foster the transparency of their actions and that inspire confidence in the profession. A distinct body of ethical norms helps archivists navigate the complex situations and issues that can arise in the course of their work.")

75 Corporate archives have traditionally been rare occurrences throughout modern history, with the notable exception of firms like Bell Labs and Xerox PARC.

76 *Dark Archive*, SAA Dictionary (updated 2020) (ebook), https://dictionary.archivists.org/entry/dark-archives.html [https://perma.cc/B39D-DSZV].

77 It may be useful to conceive of a hierarchy of degrees of archival data obfuscation indexed to the breadth of the audiences to which that data can be released. While aggregated data (the stuff of transparency reports) may be at such a distance from the content itself that it can be released publicly, it will lack the granularity needed by researchers. Heavily de-identified or less detailed data may be circulable in broad research communities, as the harms implicated by leaks are minimized, but those protections may compromise its usefulness. Lightly de-identified or masked data would be maximally useful to researchers and others, but its sensitivity would make widespread circulation less feasible, and agitate for a more restrictive access model.

78 *See, e.g.*, Noam Scheiber, *When Scholars Collaborate With Tech Companies, How Reliable Are the Findings?*, N.Y. Times, (Jul. 12, 2020), https://www.nytimes.com/2020/07/12/business/economy/uber-lyft-drivers-wages.html?referringSource=articleShare [https://perma.cc/4AET-U4K4].

79 One way to broaden the researcher base while avoiding the churn of a system premised wholly on individual permissions might be to grant access on an institutional basis. Universities could develop and administer access systems of their own with the blessing of a platform, addressing some potential criticisms around platform "cherry-picking' of researchers.

80 Kelsey Piper, *Exclusive: Google cancels AI ethics board in response to outcry*, Vox, (Apr. 4, 2019), https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board [https://perma.cc/YRC4-ZHN9].

81 Of course, transferring custodianship of archives from platforms to libraries would place a large logistical and financial burden on the latter. This burden would need to be offset by long-term funding from the platforms, foundations, or some

other appropriate source.

82 In addition to the inherent optical benefits of leadership in innovating new accountability structures, platforms may be incentivized to collaborate on Giftschrank by the systemic nature of tech's accountability and data sharing problem. In their paper, "Self-Inflicted Industry Wounds: Early Warning Signals and Pelican Gambits," Thomas Donaldson and Paul J.H. Schoemaker expound the concept of a "pelican gambit," whereby market competitors collaborate—even at their own short-term expense—to address systemic risks that threaten to destabilize the market as a whole. Thomas Donaldson and Paul J.H. Schoemaker, *Self-Inflicted Industry Wounds: Early Warning Signals and Pelican Gambits*, 55 Cal. Mgmt. Rev. 24 (2013).

83 Archivists are struggling with—and making progress on—long-term access barriers to digital content, particularly in relation to questions of transparency. *See, e.g.*, Jason R. Baron & Nathaniel Payne, *Dark Archives and Edemocracy: Strategies for Overcoming Access Barriers to the Public Record Archives of the Future*, 2017 Conference for E-Democracy and Open Government (CeDEM) 3-11 (2017).

84 *See, e.g.*, Committee on Strategies for Responsible Sharing of Clinical Trial Data, National Institute of Medicine, Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk (2015).

85 *See, e.g.*, Darren B. Taichman et al., *Data Sharing Statements for Clinical Trials—A Requirement of the International Committee of Medical Journal Editors*, New England J. of Med. (2017).

86 Lawrence Lessig, *Against Transparency*, New Republic (Oct. 9, 2009), https://newrepublic.com/article/70097/against-transparency [https://perma.cc/3TQ4-GDQ7].

87 David E. Pozen, *Transparency's Ideological Drift*, 128 Yale L. J. 100 (2018).

88 For an extensive list of COVID-19 hoaxes last updated in March 2020, see Jane Lytvynenko, *Here's A Running List Of The Latest Hoaxes Spreading About The Coronavirus*, Buzzfeed (Mar. 24, 2020), https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes [https://perma.cc/CW4X-98M9].

89 Emma Llansó, *supra* note 67.

90 Louise Matsakis, *Coronavirus Disrupts Social Media's First Line of Defense*, Wired, (Mar. 18, 2020), https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/ [https://perma.cc/9NV3-PE66].

91 Casey Newton, *How Facebook is preparing for a surge in depressed and anxious users*, Verge (Mar. 19, 2020), https://www.theverge.com/2020/3/19/21185204/facebook-coronavirus-depression-anxiety-content-moderation-mark-zuckerberg-interview [https://perma.cc/PB33-72TY].

92 Bethan John & Clea Skopeliti, *supra* note 21.

93 Kate Conger, *Twitter Had Been Drawing a Line for Months When Trump Crossed It*, N.Y. Times (May 30, 2020), https://www.nytimes.com/2020/05/30/technology/twitter-trump-dorsey.html [https://perma.cc/G4ZT-UQL8].

94 *Oversight Board Charter*, Facebook (2019), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [https://perma.cc/H3BD-6KDW].

95 Take, for example, platforms' differing models around the investigation and reporting of child exploitation and abuse imagery. Platforms that actively investigate and report on the volume of such content face optics challenges relative to those that do not, even if such imagery is pervasive industry-wide. *See* Casey Newton, *How the spread of child abuse imagery online is changing the debate over encryption*, Verge (Oct. 1, 2019), https://www.theverge.com/interface/2019/10/1/20890135/nyt-child-abuse-imagery-investigation-facebook-platforms-security-freedom [https://perma.cc/E2Y9-DKXC].

96 *See, e.g.*, Christine L. Borgman, *Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier*, 33 Berkeley Tech. L. J. 365 (2018), https://www.btlj.org/data/articles2018/vol33/33_2/Borgman_Web.pdf [https://perma.cc/54AN-XA9Q]; Sébastien Martin et al., *Open Data: Barriers, Risks, and Opportunities*, *in* 13th European Conference on eGovernment: ECEG 2013 301 (2013).

97 Directive 2016/679 of the European Parliament and of the Council of 4 May 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such

data, and repealing Directive 95/46/EC (General Data Protection Regulation) 2016 O.J. (L 127).

98 Cal. Civ. Code §1798.100

99 Jim Maddock, Robert Mason & Kate Starbird, *Using Historical Twitter Data for Research: Ethical Challenges of Tweet Deletions, in* Proceedings of CSCW Workshop on Ethics (2015).

100 A rich literature and set of best practices exist around data de-identification in the social and health sciences. See guidance documents from the U.S. Department of Health and Human Services and the U.S. Department of Education, *e.g.*, Department of Health and Human Services, G*uidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (2012), https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [https://perma.cc/H8A2-HZXT]; Department of Education, *Protecting Student Privacy: Researchers* (updated 2020), https://studentprivacy.ed.gov/audience/researchers [https://perma.cc/67B3-H574]. However, social media data may raise distinctive considerations and challenges due to its scale, its accessibility, and the personal nature of much of the data transacted via social media. For a treatment of some of these concerns, see Gabrielle Berman, James Powell & Manuel Garcia Herranz, *Ethical Considerations When Using Social Media for Evidence Generation*, UNICEF (2018), https://www.unicef-irc.org/publications/pdf/DP%202018%2001.pdf [https://perma.cc/65C5-6JUS].

101 For more discussion around the ethical challenges of using public social media content in research, *see, e.g.*, Matthew L Williams, Pete Burnap & Luke Sloan, *Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation*, 51 Sociology 1149-1168 (2017); and Annette Markham & Elizabeth Buchanan, *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee*, Ass'n of Internet Researchers 7 (2012), https://aoir.org/reports/ethics2.pdf [https://perma.cc/AM3G-PWTY].

102 One approach to this problem of legally coerced premature access to archival materials might involve the use of "time-capsule encryption," cryptographic systems designed to technologically bar access to protected content until some appointed condition has been met. *See* Jonathan Zittrain, *Time capsule crypto can help us commit our secrets to history*, The Future of the Internet Blog (Jun. 9, 2014), http://futureoftheinternet.org/2014/06/09/time-capsule-crypto-can-help-us-commit-our-secrets-to-history/ [https://perma.cc/TVP8-LYCS]. A longer paper on this topic, outlining a technical and institutional framework for implementing a system like the one described in the above blog post, will be presented by John Bowers, Jayshree Sarathy, and Jonathan Zittrain at the 2020 Digital Library Federation Forum. A draft is on file with John Bowers.

103 The Ethics of Medical Data Donation: A Pressing Issue (Luciano Floridi & Jenny Krutzinna eds., Springer 2019).

104 *About Lumen* (2020) https://www.lumendatabase.org/pages/about [https://perma.cc/3UR5-DK2D].

105 Carolyn E. Schmitt, *Shedding light on fraudulent takedown notices*, Harv. L. Today (Dec. 12, 2019), https://today.law.harvard.edu/shedding-light-on-fraudulent-takedown-notices/ [https://perma.cc/X27S-PDR4].

106 Andrea Fuller, Kirsten Grind, & Joe Palazzolo, *Google Hides News, Tricked by Fake Claims*, Wall St. J. (May 15, 2020), https://www.wsj.com/articles/google-dmca-copyright-claims-takedown-online-reputation-11589557001 [https://perma.cc/9SWP-XZUH].

About the Authors

John Bowers is a J.D. candidate at Yale Law School and an affiliate at the Berkman Klein Center for Internet & Society at Harvard University, where he worked as a Senior Research Coordinator before coming to Yale. Bowers' research interests center primarily on questions of content governance, intermediary liability, and artificial intelligence.

Elaine Sedenberg leads global research and academic engagement for Facebook's Privacy and Data Policy team and is an Affiliate at the Berkman Klein Center for Internet & Society at Harvard University. She has a Ph.D. from the Berkeley School of Information, where she completed her dissertation "Information-intensive innovation: the changing role of the private firm in the research ecosystem through the study of biosensed data." Sedenberg's research challenges the theory of linear innovation, and explores how research strategy, practice, and data policies intersect within a modern information firm. Sedenberg previously served as the Co-Director of the Center for Technology, Society & Policy (CTSP), and held a Science Policy Fellowship at the Science and Technology Policy Institute (STPI) in Washington, DC. She was awarded a prestigious National Science Foundation Graduate Research Fellowship, and also holds a B.S. in Biochemistry from the University of Texas at Austin where she graduated with highest honors.

Jonathan Zittrain is the George Bemis Professor of International Law at Harvard Law School and the Harvard Kennedy School of Government, Professor of Computer Science at the Harvard School of Engineering and Applied Sciences, Director of the Harvard Law School Library, and Co-Founder of the Berkman Klein Center for Internet & Society. His research interests include the ethics and governance of artificial intelligence; battles for control of digital property; the regulation of cryptography; new privacy frameworks for loyalty to users of online services; the roles of intermediaries within internet architecture; and the useful and unobtrusive deployment of technology in education. His book, *The Future of the Internet—And How to Stop It* (Yale University Press, 2008), predicted the end of general purpose client computing and the corresponding rise of new gatekeepers.

Acknowledgments

This paper would not have been possible without the input and feedback of many not listed on the author line. Thanks to Jennifer Allison of the Harvard Law School Library, Sam Greenspan of Bellweather, and Stephan Kellner of the Bavarian State Library for offering up their invaluable expertise throughout the research and drafting process. Thanks also to the symposium's organizers, and particularly to Katherine Glenn Bass and Amy Kapczynski, for their substantive and organizational contributions. In addition, thanks to Rob Sherman, Hershel Eisenberger, Bijan Madhani, and Alejandra Parra-Orlandoni for their feedback and perspective on concepts explored in the paper, and to Daphne Keller, who suggested the "Giftschrank" label, in light of its history, in a 2016 conversation.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Illustration: ©Erik Carter

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY