ANNEX 2

SCALING TRUST ON THE WEB

BUILDING OPEN TRUST AND SAFETY TOOLS

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

The mission of the Digital Forensic Research Lab (DFRLab) is to identify, expose, and explain disinformation where and when it occurs using open-source research; to promote objective truth as a foundation of government for and by people; to protect democratic institutions and norms from those who would seek to undermine them in the digital engagement space; to create a new model of expertise adapted for impact and real-world results; and to forge digital resilience at a time when humans are more interconnected than at any point in history, by building the world's leading hub of digital forensic analysts tracking events in governance, technology, and security.

ISBN: 978-1-61977-279-3

This report is written and published in accordance with the Atlantic Council Policy on Intellectual Independence. The authors are solely responsible for its analysis and recommendations. The Atlantic Council and its donors do not determine, nor do they necessarily endorse or advocate for, any of this report's conclusions.

Please direct inquiries to:
Atlantic Council
1030 15th Street, NW, 12th Floor
Washington, DC 20005

For more information, please visit www.AtlanticCouncil.org

June 2023

ANNEX 2

BUILDING OPEN TRUST AND SAFETY TOOLS

TABLE OF CONTENTS

INTRODUCTION

Trust and safety practices (T&S) are often misperceived as only a policy challenge for tech services to tackle, a matter of whether they have the right policies in place for managing potentially harmful content and behavior to keep users safe.[1] In reality, there is a technical implementation layer that is highly complex and is often built over time as a homegrown tooling suite and organizational structure as a service begins receiving user complaints about abuse on the platform. T&S is as much a logistics challenge as a policy challenge—a matter of facilitating effective decision-making about content and conduct, undergirded by technology.

While T&S has essentially existed as long as Internet services have, it is still maturing as a field. In recent years, organizations have begun to fill crucial gaps—for instance, by providing training and support to practitioners in developing policies and organizational structures, as well as support for organizations in developing risk-assessment frameworks.

T&S tooling is an area that remains ripe for intentional, collective investment and focus. More effective, openly available tooling—as well as more accessible best practices for development of T&S tools—could lower barriers to the development of, and competition among, a diversity of services, making it so that each organization does not need to reinvent the wheel. Moreover, it can help address what is essentially a market failure—individual services may not internalize all the social costs of harmful content and behavior, and, thus, may not invest sufficiently in socially optimal T&S.

In this paper, we consider the role open tools do, and could, play in improving T&S for a broad array of stakeholders, and where philanthropic (or other) investment might most usefully be directed. By open, we mean to be inclusive of: open-source software tools; open or pooled data; and shared-tooling solutions that may rely on proprietary software, technical resources, or services, but which are deliberately structured to be available to a wide range of platforms.

[1] We use the term trust and safety (T&S) throughout to refer to the field and practice of determining appropriate content and behavior on an online service, and managing content and conduct-related risks. The teams that carry out this work go by a variety of names; for instance, they may also be referred to as "integrity" teams. Furthermore, T&S will include product and engineering teams themselves and will intersect with other product and engineering teams in other departments. For simplicity, we simply use the umbrella term T&S.

There are opportunities at each layer of the T&S operation to support the ecosystem through open tools. Our key conclusions are as follows.

1. The logistical aspects of T&S operations appear ripe for the development of robust open tooling. Services depend on a range of tools to manage key workflows, including confirming enforcement decisions, logging and measuring decisions over time, and ensuring transparency to the public, policymakers, and others. Relevant tools include rules engines, which automate the processing and management of potentially violative content or behavior; review consoles, which provide interfaces for human review; and case-management systems, which allow services to track actions taken on individual instances of detected content and behavior.
2. Some types of detection tools provide clear opportunities for open solutions. Services use hash-matching tools to detect exact and near-exact matches of previously identified content, and some of these tools are already open source. In deploying these tools, services must consider not only their effectiveness, but also the extent to which public access to a tool's inner workings may also benefit bad actors who want to circumvent detection. Services can also benefit from best practices and toolkits that help them build classifiers that can help assess new, previously unseen content or behavior.
3. Content-specific detection tools present a complex challenge, demanding greater forethought to overall governance and institutional support. While a wide array of services may have policies against common types of content (e.g., hate speech), services' individual policies vary, no one tool will suit all, and detection tools must be updated over time. Moreover, these tools may raise complex legal questions—for instance, those related to processing of personal data. In turn, creating shared databases of violative content or content-specific classifiers raises many questions beyond simply technological design. While this is a more complex endeavor, it can provide significant utility.

More generally, ongoing maintenance, improvement, and other stewardship can shape the utility of open tools. This is particularly the case for T&S, in which tooling needs to be responsive to an ever-changing adversarial landscape. While open tools exist that tackle some of the opportunities above, they are inherently limited by the dispersed and disaggregated nature of tooling, as well as the lack of support for ongoing development. In turn, there is an opportunity to build an institutional hub for best practices that could provide technical tooling support for T&S. Such an institution could serve as a one-stop shop for practitioners who are navigating the tooling landscape, as well as contribute to development of new tools, and—critically—act as a steward of open-source tools that other actors are willing to contribute but unable to maintain in an ongoing way.

METHODOLOGY

This paper builds on workshops and interviews with about fifty T&S expert professionals, who could speak to the needs of a variety of types of organizations (companies, nonprofits, vendors) of different sizes and focuses. While this is only an initial sample of perspectives, the report synthesizes common themes to suggest a path forward that could benefit the ecosystem. We held five workshops, one of which was in person and four of which were virtual, with practitioners, experts, and stakeholders in T&S tooling. Participants included T&S practitioners from a wide variety of services, including some of the largest platforms and many smaller service providers, cutting across many different use cases (e.g., social media, commerce, real-time communications, dating, media sharing, discussion forums) and media types. We also consulted specifically with noncommercial service providers, vendors building enterprise solutions specific to T&S challenges, and academics. Following these workshops, we conducted several independent interviews with experts, and compiled desk research and data made available to us by GitHub to produce the analysis in this report. While the workshops and interview sessions were conducted under the Chatham House Rule, we have included a few quotes with permission.

THE T&S TECH-TOOLING LANDSCAPE

In this report we are focused on technologies that support T&S operations, especially as a service starts to develop and scale. Broadly speaking, T&S operations can be thought of as an iterative loop moving through four distinct phases, and then back again: detection, enforcement, measurement, and transparency.[2] While this can be a linear process, it is often more interactive, as systems must not only be improved over time but also respond in real time to actors who try to game or subvert the system.

DETECTION

Organizations identify potentially violative content or conduct based on reports received from users, or from automated tools. These tools either match a piece of content against a database of known violations or rely on advanced technology to assess the likelihood that a piece of content or given behavior is violative.

CONFIRMATION AND ENFORCEMENT

Once potentially violative content or conduct is detected, organizations must decide what action to take—for example, removing content, downranking content, or banning users. Automated rules engines help manage this process, taking immediate action in response to detection in certain cases (e.g., where automated tools are highly likely to be accurate and the potential harm is severe), or routing and prioritizing content or conduct for manual review. Software tools are also relied on for case and workflow management, providing T&S teams with an interface to track and take action.

MEASUREMENT AND INFRASTRUCTURE

Tools are used to log data about enforcement decisions and subsequently analyze them for the purposes of measuring enforcement effectiveness, audit and detection of abuse of the system, and further training of automated detection mechanisms, among other things. These data are also analyzed for internal management and regulatory compliance.

TRANSPARENCY

Organizations need tooling to message enforcement decisions back to users, facilitate appeals processes, and report on T&S operations to the public, policymakers, and other audiences.

[2] While we have used this simplified distillation, one might also think of T&S in terms of a "tech stack." A tech stack is a set of tools that serves particular purposes and is aligned to a product-development process, which can broadly be generalized to backend, midlayer, and frontend components. See e.g., Zoom's discussion of its T&S "tech stack." From this vantage point, detection, confirmation and enforcement, measurement, and transparency are the relevant goals of the "stack."

In addition to this operational loop, it is worth noting the integral role that product development as a whole can play. Product features can facilitate addressing harms, and can also proactively promote content and behavior that are understood to be additive to the community. For instance, services may develop methods for detecting and surfacing high-quality news content, or use feedback mechanisms to inform detection of other types of content. Moreover, product features may include specific controls or tools for users to manage or address harmful content or conduct. For instance, users can block or mute messages from particular users, or users might select filters that screen out certain content for themselves (while the content remains available to others on the service). These features may produce data that are then used to inform how T&S detection tools, or other parts of T&S operations, function. This broader suite of product features is not the focus of this report, although we will come back to user tools briefly in the final section.

On certain services, community-moderation systems can play a central role. These systems allow users to moderate comments, posts, and other activities within a service. They figuratively sit atop an organization's T&S operations. Examples include Discord and Reddit's models for moderating content in different channels; and Facebook and Nextdoor's models for moderating user behavior in groups. In the case of Wikipedia, effectively all content editing and moderation is managed by users themselves, and the Wikimedia Foundation's T&S function focuses on supporting the community and a specific range of abuse types (e.g., legal removal requests). To some extent, users who fill community-moderation functions face similar issues to a service provider's T&S teams at the platforms themselves: detecting violative content and making enforcement decisions. At the same time, tools made available to these moderators will also vary, and they are not our key focus here.

In evaluating opportunities for open and shared tooling solutions, we have focused on elements of the stack that are content agnostic and elements that are content specific. For example, a workflow or case-management tool for a given type of media (e.g., text, audio) may need to incorporate categories for hate speech, harassment, and many other content categories, but the most basic workflow elements are a universal part of every T&S operation in every organization and are necessary regardless of the specific category at issue. In contrast, a classifier to detect content that violates a specific rule (e.g., hate speech) will be specific to the type of content being detected.

The degree of content specificity has implications for how open a tool can ultimately be. Content-agnostic tooling is generally going to be more amenable to open development and shared solutions, for the simple reason that it is more universally applicable. While there are also a number of opportunities to open up content-specific tooling and solutions, these may face more limitations and challenges given the need for greater customization by the organizations that implement them.

HOW TOOLS ARE BUILT TODAY

Historically, when organizations had needs for T&S tooling, the build-or-buy decision was made for them by a paucity of fit-for-purpose tooling solutions on the market. In addition, services must carefully consider the privacy and security demands of T&S; for example, T&S may involve review of both public and private content and behavior, and services must carefully delineate who can access both T&S tools and resulting logs. As a result, custom solutions have been developed in house time and time again, often solving variations on a more generalized problem seen before by other services.

In recent years, a number of vendors—some start-ups, some larger enterprises—have emerged to provide tooling solutions to T&S problems. While these companies signal the emergence of market solutions for tooling needs, T&S professionals who we interviewed voiced concerns about dependence on a limited set of firms, particularly as start-up vendors have tended to eventually be acquired and brought in house by larger companies. For instance, in 2018, Twitter acquired anti-abuse provider Smyte, and the tool was taken off the market. Our initial scan of market data suggests that as many as one-third of all T&S start-ups in the past decade were acquired by other companies. In many of these cases, the former customers of the start-up may lose access to an external tool on which they had relied, and must then reinvent the wheel with in-house development or by finding another vendor.

Open tools exist today, and are put to use to varying degrees. In our interviews, practitioners suggested that open tools might provide a starting point from which service providers can build, but that there is not a go-to set of tools on which many could readily depend. Instead, there is a vast array of different options available, and it is challenging for services to determine the suitability of any given tool.

A core challenge in providing open-source tools is the need for ongoing maintenance and customization. Nothing in T&S is "set and forget"—tools need constant maintenance to stay relevant to the evolving threat landscape online. Practitioners with whom we spoke pointed to the risk of relying on an open-source solution if it is not well supported, or if there is a chance the originating organization may cut support in the future. As we will discuss further below, these challenges with the open-source tooling ecosystem today present an opportunity for focused investment and support.

BUILDING SHARED TECHNOLOGY TOOLING

In the sections that follow, we describe the elements of the T&S operational loop, and the varying degree to which these elements are amenable to open-source tooling and shared solutions. Each operational element has different component parts, some of which are content specific and others of which are content agnostic.

DETECTION

The most basic mechanism through which platforms detect potentially harmful content and behavior is a user-initiated action; most platforms provide some mechanism, often a flag or similar button to "report a violation," for users to report violative content or harmful behavior. While this functionality is common, it ultimately needs to be tailored and customized to fit a service. Best practices and reference implementations could still usefully inform how services build their own systems, and, as we discuss later, ensuring these user-initiated actions are routed to the appropriate teams is an element of workflow management in which open tooling could be helpful.

Similarly, some services use "trusted partner" or "trusted flagger" mechanisms that allow nongovernmental organizations (NGOs), governments, and other organizations to access more sophisticated reporting tools than are available to the average user. These programs often route trusted partners' reports with different prioritization, as organizations rely on partners for language and cultural or other expertise in evaluating content. On the one hand, these tools are also customized to integrate into a given product or service, and, thus, are not prime candidates for open tool building. On the other hand, these programs are increasingly becoming required of some platforms by regulation, and may, therefore, become standardized to varying degrees. What is more, participants in our interviews noted that these systems were often not tailored for a diversity of partners, including organizations from the Majority World.

In addition to detecting abuse reactively via user-initiated action, services also deploy automated, proactive detection tools for violative content and behavior. The scale of these efforts can vary with the purpose and scale of the organization, but they generally exist at least in some form (for example, to address spam).

In particular, we spoke with experts about two common tools: hash matching for exact and near-exact content matches; and classifiers built with machine learning or other AI tools to assess new, previously unseen content for potential violations.

HASH MATCHING: DETECTING EXACT OR NEAR-EXACT MATCHES OF VIOLATIVE CONTENT

Hash matching is a method used to detect content that is exactly the same or nearly the same as previously identified content. Exact hash matching identifies exact matches to known violative content. "Fuzzy matches" and perceptual hashes are used to find content that is nearly identical to previously identified content, but different enough to be missed by exact hash matching.[3]

Fundamentally, hash matching is a content-agnostic method for creating and identifying hashes of content and matching that content to known violations, and both exact and "fuzzy match" methods can be, and have been, open source. For instance, common hash functions are readily available as open-source libraries. Companies have also contributed code to improve methods. For instance, Facebook has opened its photo and video hash-matching tools, PDQ and TMK+PDQF, respectively, as well as a more comprehensive Hasher-Matcher-Actioner tool that facilitates labeling, matching, and actioning violative content.

Hash functions are designed to impede reconstructing the original content simply from the resulting hash. While some argue that open tools may be more vulnerable to adversarial attacks, others suggest that "security through obscurity" is not effective here, and that the benefits of open-source contributors overall support the effectiveness of these tools.

Of course, effective hash matching depends on having a database of hashes. Typically, a service maintains a database of hashes based on content it has already addressed. In addition, services might rely on shared databases of hashes, matching content on their service against a database of previously identified content from elsewhere.

The most common instance of this approach exists with child sexual-abuse material (CSAM). For instance, the National Center for Missing and Exploited Children (NCMEC), International Watch Foundation (IWF), and other similar organizations maintain databases of CSAM (pursuant to legal guidelines), and services can then check against these databases to take action against violative content on their systems.[4] In many cases, services access shared databases by using hash-matching tools that are offered as a centralized, though broadly available, service in the context of addressing CSAM. Microsoft, Google, and Cloudflare have all built CSAM hash-matching engines that check content against databases of known CSAM, and which they make available to other platforms via API or, as in Cloudflare's case, customers.

In recent years, databases of hashes have also been developed for shared use in other contexts. Most notably, the Global Internet Forum to Counter Terrorism (GIFCT) has developed a cross-platform, shared hash database of terrorist and violent extremist content (TVEC) that member organizations can use to identify content on their platforms. Unlike CSAM, TVEC does not have a universal definition and is not universally illegal, and each organization can decide to act on this content in different ways.

A similar institutional solution has been developed for nonconsensual intimate-image abuse. StopNCII.org is a hash database operated by the UK Revenue Porn Helpline, a nonprofit organization in the United Kingdom (UK). Adults are able to generate a hash of intimate imagery that was created without their consent; this hash is created locally on a user's device, and then only the hash is sent to StopNCII.org for inclusion in the database. Participating companies will run their content against this database to detect and remove matches.

[2] Here, we use hashing in the way the Office of Communications (OfCom) does in its extensive report: "Hashing is an umbrella term for techniques to create fingerprints of files on a computer system." See also Hany Farid's "An Overview of Perceptual Hashing."

[3] For a deeper analysis of this topic, see *Annex 3: Respecting Children as Rights Holders*.

These content-specific hash databases, and associated hash-matching tools, are another area for possible investment, but they raise much more complexity. Investment in additional hash databases and matching solutions needs to be regarded as an institutional challenge—just as much as, if not more than, a tooling challenge—engendering the trust of a wide range of stakeholders.

On the one hand, practitioners noted the utility of such tools, particularly in the case of CSAM, where the harm is severe and the content is universally illegal regardless of context.[5] On the other hand, the utility of these databases depends on strong, trustworthy governance regarding their contents. Concerns about bad actors using the database to reverse engineer and access harmful content, or accessing the hash function to circumvent it, may encumber the full openness of available tools. There is also a risk that data will be improperly included in the database, and, as a result, that organizations will be overly restrictive in removing content.

A related concern with investing in developing broadly shared databases is they would facilitate the development of "content cartels," even if done unintentionally. Because these systems can be expensive to develop and maintain, the concern is that the entire ecosystem would default to the easiest and most available, creating a de facto standard.[6]

Along with detecting violative content, another detection challenge that many T&S professionals cited is the need to share intelligence across platforms and receive intelligence from trusted partners. This is particularly the case for detection of behavioral patterns that contribute to harmful content online. For example, identifying and enforcing against disinformation often requires identifying clusters of fake accounts that are found to be acting in concert, or other forms of coordinated inauthentic behavior. Today, some services have developed arrangements to share information with industry peers, but doing this in a way that respects privacy is challenging. Regardless, creating an institution to facilitate threat sharing would raise many of the complexities of the hash databases noted above. Again, the challenge here is as much institutional as it is technological.

CLASSIFIERS AND OTHER AUTOMATED ASSESSMENT OF PREVIOUSLY UNSEEN CONTENT

Hashing solutions help with identifying exact or near-exact matches to previously seen content. But what about new content? Services use a variety of approaches, including automated systems that monitor for the characteristics of bots used to manipulate platforms; text-analysis tools; and a variety of approaches based around machine learning and artificial intelligence (AI). For instance, through machine learning and other AI techniques, services create and deploy "classifiers" that automatically assess new content and behavior to assign scores that reflect the likelihood of violations. These tools are a linchpin tool for detection of abuse, automating a first-pass evaluation before passing things into a queue for human review.

Practitioners noted that the T&S ecosystem would benefit from best practices and toolkits that facilitate the development and evaluation of classifiers and other detection tools. Services could benefit from guides and reference implementations that assist in the process, and approaches for measuring and evaluating efficacy. For example, Google has provided a reference implementation of how to use the open-source TensorFlow platform for creating content-moderation tools. While the implementation uses an example classifier created for detection of toxic content, anyone can take this model, install it, and run it on their own dataset to develop a classifier that is fit for purpose.

[5] In addition, in the context of CSAM, there are restrictions on even possessing the content. As such, there are some benefits to relying on a third party that has the requisite permission to operate a database of this sort.

[6] In addition, to the extent the tools rely on passing a service's content to a third party for analysis and matching, it raises potential competitive and privacy concerns for the originating provider. There may be ways to address this concern by creating a hash at the client side and then passing that to the server. For instance, Microsoft has begun trialing such a system for its PhotoDNA system. See: Microsoft Moderator Service API Documentation, "Match Edge Hash."

Building open, content-specific classifiers is also possible, but raises institutional challenges as much as technological ones. A wide variety of open classifier tools already exist.

- An initial analysis of open-source code repositories on Github, based on only a few dozen keyword topics, found more than five hundred libraries related to content classifications; a robust analysis would surely find multiples more.
- Hugging Face has cultivated a developer community building open-AI models, and it features datasets and detection tools for different types of content.
- Social media platform Bumble released a tool for "lewd photo detection."
- Jigsaw, an Alphabet company, has contributed its conversational-AI-moderator code to Github, which can be used to detect toxic comments; Jigsaw has also released this as Perspective API that others can use.
- Startups like Unitary are also contributing classifiers to the market, and are committed to building more with open source.

Despite the appearance of a robust set of open classifiers, practitioners suggested that, today, these tools are of only limited utility. To begin with, it is challenging to even know what is available—there is no trusted source or compilation of all the open tools that exist. Even when organizations have deployed classifiers that were originally developed externally—in an open-source fashion, or in cases where classifiers have been brought in via acquisition of start-ups—T&S professionals must heavily customize the tools. Even with a tool available as open code, it can be difficult to evaluate and compare its performance, including how it was trained. For example, we heard in our research that even with a seemingly widely used classifier, such as Yahoo!'s Open NSFW tool (which focuses solely on detecting pornography), many organizations that use it take it as a baseline input on top of which they build further customization for their platforms' specific needs. Expert Adelin Cai, who is a co-founder of the Trust & Safety Professional Association and worked on T&S teams across a number of companies, has seen multiple instances of services paying for third parties to perform initial screening of content on the platforms, as well as to provide lists of keywords related to potentially harmful content. Nevertheless, the services still need to do in-house customization due to the uniqueness in how each organization would use the output, so "wouldn't it be great if there were open options for those organizations that just need somewhere to start."

Classifiers also need active stewardship to be effective. Like spam filters or other systems intended to identify or screen certain types of content, classifiers operate in an environment where adversaries are continuously working to game the system and push their violative content through. As a result, any investment in this tooling needs to account for ongoing governance of the tool, ideally embedded inside an organization tasked with its ongoing upkeep.

Moreover, deployment of third-party classifiers raises issues similar to those of the hash databases discussed above. While access to code can help support a greater degree of transparency and accountability, it is insufficient to fully understand how the classifier operates. To deploy a classifier, a service provider has to trust the data and that the training of the tool was done in a way that aligns with its values and policies—and that may not be the case. In every case, classifiers may carry with them bias, unfairness, or inattention to a variety of other contextual factors. Even something as seemingly mundane as a profanity classifier could be trained in a way that is more or less attuned to certain vulnerable communities.

Nevertheless, participants repeatedly pointed to the limited geographic reach of existing classifiers as a key market failure, and an area in which investment in open and shared classifiers could address gaps in a meaningful way. This may be the case with languages and communities in the Majority World, for example. Data availability in certain languages may be highly limited, and practitioners noted that companies whose primary revenue-driving markets are English language and culturally Western are unlikely to invest in building high-quality classifiers for other markets and languages.

Another opportunity for increased investment is the contribution of datasets on which classifiers can be trained and evaluated. Particularly for small organizations, building the underlying dataset can be a lot of work, as it requires both the underlying content and the human and computational investment in labeling the dataset. While a user of these sets must still be discerning about the underlying content before using it as an input in the model, these could still provide useful starting points.

ENFORCEMENT

After detection of harmful content, T&S teams must confirm an assessment of the content and implement an enforcement decision. Many tools are currently used to do this both manually and automatically, and some of them might be good candidates for open and shared development. In this area, we found relatively little when it comes to existing, open tools tailored for T&S; however, practitioners suggested that these aspects held promise for shared solutions.

RULES ENGINES

As the number of content flags scales with a platform's growth, many T&S teams quickly find themselves inundated beyond what they can manually react to in a timely fashion. Rules engines are built to provide a first run of automated processing on high volumes of content, and several experts spoke of them as critical tools in the T&S toolkit. These engines automate some enforcement decisions, but primarily route and prioritize decisions for content and conduct that has been classified already.

While different services will create different rules, rules engines themselves are a general need, and are a potential opportunity for building open tools. Alex Feerst, a leading expert and former general counsel and head of trust and safety at Medium, was among the many practitioners to note how rules engines were a critical part of the "core plumbing" of T&S operations:

> "Every company needs to think about the logistics of detecting and reviewing content and conduct. They need to ensure both automated enforcement and human review are deployed well. Triage, prioritization, and routing are key parts of any well-crafted system. To some extent, each service will probably need customization. Yes, each service has its own structure and focus that will guide the design of a functional system. But they are also not so unique that we need to reinvent the wheel for each one. There's an equilibrium somewhere between understanding the idiosyncrasies of each product or community, and drawing on shared concepts and approaches we can apply more broadly."

QUEUEING AND WORKFLOW-MANAGEMENT TOOLS

Based on our interviews, most T&S teams start out on the most generalizable tooling solutions for workflow-management and queueing needs: some combination of providers like Zendesk and Jira, or similar SaaS (software as a service) solutions for ticketing and customer support. These tools are used to solve workflow needs, but more tailored solutions would help, particularly as platforms scale; a hacked-together workflow needs to integrate new tools or features to accommodate changing organizational structures, product needs, and review features.

This integration challenge was cited as a key tooling challenge for which open solutions could play a role. Open solutions for workflow management may look like a fit-for-purpose ticketing and enforcement interface, and may augment existing work to simplify prioritization and help organizations tackle queueing challenges for specific types of content.

When it comes to the interfaces used to review particular content and behavior, practitioners called out the importance of supporting reviewers' well-being. For instance, to mitigate the risk of exposing reviewers to grossly violent or abhorrent imagery, content-moderation systems may grayscale and blur images, allowing a reviewer to determine whether action can be taken without any further detail and context.

In addition, practitioners noted that regulatory requirements are increasingly directly relevant to case management. For instance, the Digital Services Act specifies how hosting services must provide users with a "statement of reasons" regarding removed content. To the extent these sorts of requirements drive some measure of standardization in T&S workflows, open solutions may also be helpful in providing off-the-shelf support for a variety of services.

MEASUREMENT

When violative content or conduct is detected and action is taken, that information is then fed back as a data point to further train automated detection mechanisms, build profiles of abusive behavior, and evaluate T&S' operations performance.

Services' data architecture may vary a great deal, and, thus, participants in our interviews suggested it would be difficult to generalize such a system. However, it could be possible to develop basic, open logging tools that build upon the enforcement infrastructure, labeling actions taken on various content.

Similarly, while different services track different trends, all need the capacity to produce at-a-glance analysis of their enforcement efforts. A general, open tool could be created to facilitate some of the most common elements that T&S teams might want to track across different types of content. Furthermore, such tools could incorporate common, best-practice metrics that support quality assurance; currently, public versions of these frameworks do not appear to exist, according to the practitioners to whom we spoke. Along with tracking metrics around violative content, practitioners also called out the opportunity for common frameworks for metrics around positive or "prosocial" uses of a tool (e.g., sharing of content meant to reduce intergroup hostility) and interaction with features designed to support such uses (e.g., features that remind people to be civil, and prompt them to consider the content their posting).

TRANSPARENCY

While much of their data will remain internal, services make data transparent to users, researchers, the public, and, increasingly, regulators. This transparency has associated tools needs.

Services build tools to report to a user that their content has been removed, and sometimes to provide appeals processes. They also might build dashboards that allow users to track content they have flagged, so that they can see whether it has been acted upon. These are highly customized to a service.

Services also build external transparency reports, reporting aggregate statistics and other information about their T&S processes to both the public and regulators. Aggregate transparency reports vary, too, in final presentation. However, services generally look to report common items like the number of pieces of content removed in a given category of content. Moreover, to the extent that legal requirements drive some amount of baseline convergence in reporting needs, a shared, open tool could provide a fit.

BUILDING A HUB FOR BEST PRACTICES

How might work across these areas be advanced? Practitioners did not have a unified view, but they pointed in a few directions. The overarching challenge is that there is no one-stop shop to consult to understand the tools that are available or to get advice on what considerations to keep in mind when implementing them. The creation of a single hub to drive a critical set of activities could be prudent.

A starting point for this work might be to simply compile and curate lists of the existing tools up and down the technology stack. Each tool might be described in a single place, detailing and benchmarking its capabilities, dependencies, and limitations, as well as what is required to implement it. In the absence of any organized effort to do this, some academic researchers have attempted to compile datasets, classifiers, and machine-learning (ML) tooling examples, but these efforts are nontrivial to identify or navigate.

We also heard from a number of T&S professionals that their teams and organizations would be interested in contributing open-source code for tools they have built in house. But many of these tools are not well suited to simply live on Github or another third-party site; rather, they require a responsible steward to maintain them over time, adapt them to specific deployments and communities, and keep them up to date with the current adversarial environment. For example, Discord wants to release an open-source version of its in-house rules engine at some point in the future, but hopes that a steward can be identified to maintain and support the tool to assist other organizations that choose to deploy it. Effective development for this and other open tools will require an organization staffed with technical experts, as well as T&S domain experts.

The opportunity to compile what already exists and support future contributions of open-source code speaks to the need for an institutionalized technical capability focused on T&S. Given optimal technical talent and organizational support, there is significant opportunity to build and contribute open-source and shared code and tools directly. This might take the form of releasing reference implementations of classifiers or other tools, or building and maintaining open-source tools, such as a workflow-management engine.

LOOKING OVER THE HORIZON

T&S is a dynamic field, and practitioners also encouraged thinking about how open tools might intersect with issues that are on, or just over, the horizon. We highlight three areas below.

ARTIFICIAL INTELLIGENCE

Over the last decade, the use of content classifiers built with machine learning and other artificial-intelligence tools dramatically changed the ability for services to identify new, potentially harmful content at scale. One of the most provocative comments we heard in our research was that we may be at another pivot point, as generative AI reshapes operational T&S over the coming years.

Researchers and services are already beginning to deploy generative AI systems to help at different parts of the T&S process.

- Microsoft researchers recently demonstrated how they used large-language models (LLM) to synthetically create a dataset of hate speech, which was then labeled and used to train a classifier.

[7] Some vendors, such as Active Fence, have begun to provide templates for transparency reporting with these regulatory requirements in mind.

- Cohere.ai posits that LLMs will allow classifiers to be developed from much smaller datasets. Rather than requiring people to label myriad pieces of data to train classifiers, an LLM can use a small, labeled set to then label other data itself.
- OpenAI used GPT-4 to develop classifiers to identify harmful outputs of its system.
- Our interviews suggested that such systems could also be set up to take enforcement actions and provide an explanation to both the system operator and the user. The system could also be used to work in the reverse direction; that is, from a set of enforcement decisions, the system could be asked to distill what relevant rules apply to them.

We cannot evaluate how far and how fast these tools will impact T&S, but investment in open tooling for T&S should be attuned to these changes. Experts with whom we spoke emphasized that there will be a growing demand for tooling that allows people to use LLMs to rapidly spin up contextually appropriate T&S operations, including rules, systems, and classifiers. In other words, the AI may exist and be capable of significant impact on T&S operations, but tooling will be required to allow people to optimally use the AI for that purpose.

AI has also gained widespread attention for the role it plays in increased capabilities for anyone to generate synthetic content at scale. This growth of synthetic content has raised questions around so-called "deepfakes" and how to detect them at scale. More generally, people can synthesize harmful content of all stripes. As noted above, datasets to train detection tools, reference implementations, and open classifiers could prove useful here.

EVOLVING SERVICE TYPES

T&S will need to evolve to adapt to the challenges of a number of new service types, including real-time communication, metaverse technologies, and the decentralized social web.

EXTENDED REALITY (XR) AND THE METAVERSE TECHNOLOGIES

XR technologies encompass both virtual-reality environments and augmented-reality tools that overlay digital objects on the physical world. These technologies have been most recently associated with the idea of a "metaverse" in which more of our Internet-enabled experience is done in real-time, embodied communication. On the one hand, real-time environments are nothing new to T&S. On the other hand, they have not been a predominant form of interaction. What's more, the types of media objects involved (e.g., three-dimensional renderings) can require different detection tooling, and the behavioral and interactive elements may lead to different harms (e.g., simulated physical or sexual assault) that require different interventions. While the overall workflow of T&S may remain consistent, it might also respond to new signals based on the sensors embedded in the devices, raising new T&S and privacy issues.

One key way in which XR may demand a different focus for T&S tooling is the sheer variety of formats, spaces, modes of communication and interaction, group sizes, and environments that users may encounter. Community-moderation tools are already relied upon by services such as Discord, Reddit, Nextdoor, and Facebook, and these types of tools may become even more important in XR to accommodate the range of group dynamics that may unfold. XR may also heighten the need for user-specific T&S tools that would allow a user to filter certain conduct or content out of their experience. For example, Meta has incorporated a feature that garbles the voices of strangers on its Horizon Worlds service, so that talking to and engaging with strangers is something that a user can choose to avoid.

DECENTRALIZATION OF SERVICES AND USER CONTROLS

As we noted at the outset, T&S operations are not homogenous. While it is easy to think strictly in terms of centralized operations run by a service provider, there are services that are largely moderated by users themselves. What is more, in part due to a backlash against centralized control of online platforms, there are an increasing number of decentralized services evolving.

Mastodon is one example that illustrates some of the T&S challenges of the decentralized web. Unlike at Twitter or Facebook, there is no central Mastodon T&S team that moderates content across the service. In fact, it is impossible for such a team to exist, as each Mastodon server instance is responsible for moderating the content hosted on that server. At present, the tools to do so are relatively rudimentary, and many of the data signals that T&S operations tend to use with centralized services (e.g., access patterns) are not necessarily usable on these services.

This is not simply an issue for Mastodon. For instance, the InterPlanetary File System, which facilitates distributed file storage, has worked to create an optional hash list for storage operators to facilitate blocking.

Participants in our interviews noted that the solutions for decentralized services may not simply be carbon copies of what works elsewhere. In particular, to the extent that the operators of individual nodes in the network are hobbyists, they may need very different sorts of tooling.

Relatedly, more decentralized architectures may open up opportunities for new sorts of user controls created by a range of parties. For instance, the Bluesky social media service is being designed so users can use different clients; in addition, developers can create different algorithmic feeds and users can select from them. It also allows people to create moderation labels that developers and users can use to support more customized content-moderation choices. To be clear, user controls do not necessarily depend on decentralized architectures for services; for instance, the Block Party is a third-party tool for Twitter users, helping them more easily filter out possible harassment. The point here is simply that new, more decentralized service architectures may expand these opportunities.

AGE ASSURANCE

Services may engage in verification processes for users for a variety of reasons, and age assurance is a subset of user-verification issues, referring to methods services can use to estimate a user's age with varying levels of certainty. While legal requirements around age assurance remain controversial, they are increasingly common, and service providers are having to reckon with how they will adapt. Determining a user's age with more granularity may trigger requirements and opportunities to apply certain safety measures; at the same time, collecting data necessary to make that determination raises a host of privacy and security questions.

Several experts in our interviews pointed to the potential benefit of open solutions in this area. We heard three key areas in which open solutions could be helpful.

Interoperability: Service providers and users could benefit from an organizational or technological solution that creates interoperability, so that a user who has their age assured via one service does not need to repeat the process at other sites. Relevant efforts are under way in Europe to create a shared solution for age verification that effectively allows users to reuse an age check across participating providers in a privacy-preserving way.

Age-inference models: One age-assurance method is through inference based on data a service already has. Just as with content classifiers, one can imagine tools that estimate age based on a common set of inputs. For instance, researchers created an open-source tool to infer demographics based on public Twitter data, and one could imagine a broader classifier for public text on social media that estimates age. Along with raising similar challenges as classifiers generally, the challenge here is that many attempts at estimation will rely on private data, and solutions that work for one provider's data architecture may not be easily transferable.

Age inference can also be done via tools that analyze a person's face. Open-source tools to do this already exist. While robust open solutions could help lower the cost of this method, it would not address the fundamental privacy challenge involved in collecting that biometric information.

Nonprofit clearinghouse: A much more ambitious approach would create an entity that would be entrusted with collecting relevant age-assurance information from a user. This would not eliminate the privacy and security concerns with collecting user information for age assurance. However, it would attempt to address them through an intermediary that is noncommercial and entrusted to the public good, separate from the government or any companies. It could necessarily require significant technical expertise to get this right—but, even more so than with some of the concepts above, the institutional challenge here is the tip of the spear.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This report was authored by Proteus Strategies, a boutique policy consulting firm based in San Francisco, California, founded by Derek Slater and Betsy Masiello. Collectively, the Proteus team has more than thirty years of experience managing communications and public policy efforts on issues such as online safety, data privacy and security, and open infrastructure from inside leading technology companies. This work was supported by the William and Flora Hewlett Foundation. The authors are grateful to the members of the Atlantic Council Task Force for a Trustworthy Future Web for providing their feedback and perspectives.

APPENDIX

OPEN AND SHARED T&S TOOLS REFERENCES

The following is a compilation of open or shared tools referenced above. As noted, there are a wide variety of open tools available, so this is not comprehensive.

HASH-MATCHING CAPABILITIES

- Facebook's PDQ, TMK+PDWF, and Hasher-Matcher-Actioner (GitHub repo) .
- Other referenced perceptual hashing tools like pHash are widely available, implemented in different forms—see, for example, Github repos.
- For CSAM, matching tools include
  - Microsoft PhotoDNA;
  - Google CSAI Match; and
  - Cloudflare CSAM scanning tool.

HASH DATABASES

- CSAM databases are managed by groups like NCMEC, Internet Watch Foundation, Thorn (its Safer Database), Interpol, the Dutch Expertise Bureau Online Kindermisbruik (EOKM), and the Canadian Centre for Child Protection (Project Arachnid).
- GIFCT operates a hash database of violent extremist content.
- StopNCII.org's database pertains to nonconsensual intimate imagery (aka "revenge porn").

CLASSIFIERS

- Google Content Safety API—focused on child-abuse material.
- Bumble's Private Detector.
- Jigsaw PerspectiveAPI and ConversationAI-Moderator tool.
- Unitary Detoxify.
- Yahoo Open NSFW.
- GitHub features many relevant libraries—see, for example, pages for hate-speech detection tools and violence detection.
- GitHub and Hugging Face similarly have many relevant models and datasets—see, for example, a search for "toxic" to find relevant detection tools on Hugging Face.
- OpenAI Moderations endpoint.

REFERENCE IMPLEMENTATION AND GUIDES FOR ML/AI DEVELOPING TOOLS

- As one example, TensorFlow published its guide to use for content moderation, as one among many platforms.
- Various GitHub libraries provide overall guidance—see, for example, "content-moderation-deep-learning."

OPEN DATASETS

- Both Github and Hugging Face contain relevant datasets for content detection—see, for example, datasets related to toxic content.
- Google contributed a deepfake dataset.
- Ubisoft and Riot announced a collaboration to build a shared dataset.

ENFORCEMENT

Rules engines, queueing, and workflow-management tools are available in various open forms, but we did not find items tailored for T&S.

- GitHub has a number of rules engines, but none tailored to T&S.
- Open-source tools like osTicket have general-purpose ticketing and case-management capabilities, much like Zendesk, Jira, and similar tools.