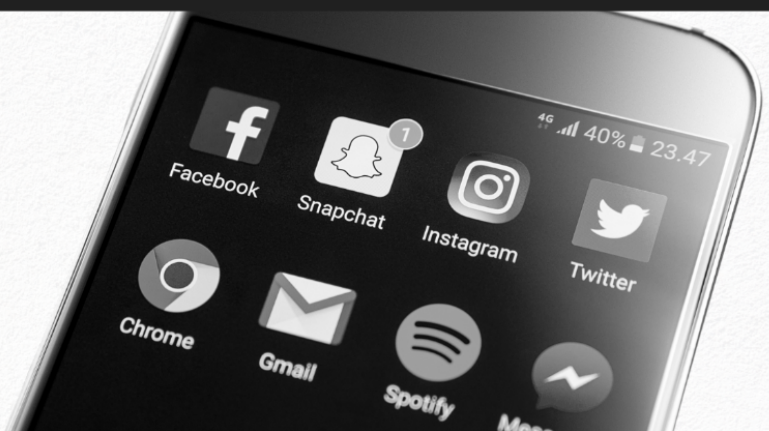
CCDH Center for Countering Digital Hate

# STAR FRAMEWORK

## A Global Standard for Regulating Social Media

SAFETY BY DES
TRANSPARE
ACCOUNT
REGIONS



# STAR Framework

### A Global Standard for Regulating Social Media

## Introduction from the Chief Executive, Imran Ahmed

A handful of companies, owned by a small coterie of Big Tech billionaires, dominate Internet content. This elite owns the technology that connects 4.5 billion people around the world and creates a platform on which individuals can share information, make new relationships, create communities, develop their brands, and transact business. Platforms produce little content themselves, but have produced business models in which they monetize the content produced by billions of people by selling both users and data on the psychology of those users to those seeking to sell their own products, services, brands and ideas.

The communities on these online platforms, their behaviors and beliefs, and the values which emerge from those spaces increasingly touch every aspect of offline society. The tech companies and their executives know the real harms their products can cause, which one would expect would lead to their curation of these environments, but the imperative for growth and acquisition of market share of eyeballs to sell to advertisers is their only concern. This was most pithily explained by the Chief Technology Officer of Meta, when he wrote in an internal memo:

> *"So we connect more people. That can be bad if they make it negative. Maybe it costs a life by exposing someone to bullies.*
>
> *"Maybe someone dies in a terrorist attack co-ordinated on our tools. And still we connect people.*
>
> *"The ugly truth is that we believe in connecting people so deeply that anything that allows us to connect more people more often is *de facto* good. It is perhaps the only area where the metrics do tell the true story as far as we are concerned."*

For many years, this elite of billionaire owners have postured themselves to operate

under the utopian charter myth of neutrality and virtuous contribution to the growth of human understanding through social media. At its core, their proposition is an old-fashioned advertising business that reshapes cheaply-acquired content by promoting the most salacious, titillating, controversial, and therefore "engaging" content. This was designed to stave off the moment that regulators might turn their eye to this industry. 'Social media' is not a synonym for the Internet or even for technology, and yet, by hiding behind the techno-utopian halo of online technological innovation, they have both hidden the banal atavism of their core business models and avoided real scrutiny of the harms they cause. Surely, many opine, this is inadvertent or unavoidable. In fact, neither of these statements are true.

The laws that seek to regulate this enormous industry which directly affects billions of people were, for the most part, created before social media companies existed. In the United States, they were codified in Section 230 of the Communications Decency Act 1996 which sought to protect bulletin boards and newspaper comments sections from third-party liability in order to foster innovation and growth for a fledgling industry. This led to decades of regulatory ambivalence and the international community adopting a 'hands-off', or at best, an individual content-based approach, to regulating online harm in some jurisdictions, with technology companies seen as neutral actors in this environment. Tech companies were encouraged - through this permissive regulatory environment that functions without checks and balances - to adopt aggressive profit-driven business strategies that follow what Mark Zuckerberg described as a "move fast and break things" maxim, as outlined in his 2012 letter to investors.

Things are, indeed, broken. Through our work at the Center for Digital Hate (CCDH), we have developed a deep understanding of the online harm landscape. Since 2016, we have researched the rise of online hate and disinformation and have shown that nefarious actors are able to easily exploit digital platforms and search engines that promote and profit from their content. CCDH has studied the way anti-vaccine extremists, hate actors, climate change deniers, and misogynists weaponize platforms to spread lies and attack marginalized groups. Through our work, we have seen the depth and breadth of harm that tech companies profit from on a daily basis, including:

- **Hate and Extremism**: including but not limited to racism, hate content targeting women, the LGBTQ+ community, and faith communities (e.g. anti-Jewish hate and anti-Muslim hate); and

- **Mis/Disinformation** on critical issues like COVID-19, climate change, and elections.

What has remained consistent, across all types of harmful content, is the absence of proper transparency and the failure of platforms and search engines to act. Our research and advocacy work shows repeated failures by social media companies to take action on harmful content or the actors/networks who spread it. We have demonstrated how the companies' algorithms - with a systematic bias towards hate and misinformation - have had a damaging impact on our information ecosystem. The failure of social media companies to act on known harmful content connected with extremism, terrorism, racism, misogyny and online hate is a violation of their own terms and conditions, the pledges they make in the media and to governments, and the basic duty they have to their users to have a right to exist safely online and in their communities. This failure to act is the reality of the status quo of self-regulation. Self-regulation means no regulation.

The status quo cannot stand.  It has a damaging impact on individuals, communities and our democracies. CCDH research has evidenced the need for legislation that addresses the fundamental business models and therefore behavior of the platforms that profit from the spread of misinformation, disinformation, conspiracy theories, and online hate by bad actors and by their own systems. CCDH has advised the UN, the UK, the US, and other governments on disinformation, violent extremism, and how conspiracy theories can overwhelm fact-checking countermeasures and cause considerable real-world harm.

Following the CCDH Global Summit in May 2022, we saw a need to develop a values and research-driven framework to support global efforts to regulate social media and search engine companies. In this document, we have set out core elements of the *STAR Framework* with explanations and examples from our research. Through the *STAR Framework,* we aim to establish key global standards for social media reform, to ensure effectiveness, connectedness, and consistency for a sector which impacts people globally.

The impact of online harm is real - on people, communities and democracy. We cannot continue on the current trajectory where bad actors opportunistically weaponize and poison the information ecosystem, which itself functions through a broken business model from Big Tech, to drive offline harm. We need to reset our relationship with technology companies and collectively legislate to address the systems that amplify hate and dangerous misinformation around the globe. The *STAR Framework* draws on the most important elements for achieving this: Safety by Design, Transparency, Accountability and Responsibility.

## Core Elements of the STAR Framework

| | |
|---|---|
| **S** | **Safety by Design**: Safety by design means that technology companies need to be proactive at the front end to ensure that their products and services are safe for the public, particularly minors. Safety by design principles adopt a preventative systems approach to harm. This includes embedding safety considerations through risk assessments and decisions when designing, implementing, and amending products and services. Safety by design is the basic consumer standard that we expect from companies in other sectors. |
| **T** | **Transparency**: There are three key areas where transparency is desperately needed and should be prioritized:<br>● Algorithms;<br>● Rules enforcement; and<br>● Economics, specifically related to advertising. |
| **A** | **Accountability to democratic and independent bodies**: Regulation is most effective where there are accountability systems in place for statutory duties and harm caused, particularly where there is a risk of inaction because of profit motives and commercial factors. Frequently, accountability systems include an enforcement and independent pathway for challenging decisions or omissions. |
| **R** | **Responsibility for companies and their senior executives:** The final element of the STAR Framework is responsibility - both social media and search engine companies and their senior executives that are responsible for implementing duties under a legislative framework. Responsibility means consequences for actions and omissions that lead to harm. A dual approach - targeting both companies and their senior executives - is a common intervention strategy for changing corporate behavior. |

# Safety by Design

## What does it involve?

Safety by design involves deliberate and proactive choices by technology companies about the design, implementation, and execution of their products and services, so that they are safe for the public. In practice, safety by design means that:

- Tech companies have tested their product and incorporated effective safety features before it is available to the public;
- Instead of tolerating or amplifying harmful content, companies have designed interventions, systems, and processes to remove it from the platform and/or minimize its spread to a broader audience, including children;
- There are accessible, effective and responsive reporting pathways built into the product to ensure that any breaches are able to be dealt with in a way that minimizes harm; and
- The company has designed ways to identify, evaluate, and modify the impact of their products and services on an ongoing basis, including when they make significant changes to those products and services and in response to reported, identified and emerging harm.

A legislative framework needs to have the right incentives in place so that safety by design is the default position. There are different ways to achieve this. For example, in the UK's Online Safety Bill there is a duty of care model coupled with significant legal liability and financial penalties for failure to comply. In the US, opening the doors to litigation through a modified version of the liability immunity in Section 230 of the Communications Decency Act 1996 may also achieve this goal.

## Why is it needed?

Principles of safety by design should apply to both new and emerging products and services. Despite promises made by Big Tech companies and their executives, safety is clearly not being prioritized or implemented to a sufficient degree. For example, Chief Executive Mark Zuckerburg and President of Global Affairs Nick Clegg, at the launch of Metaverse promised that *"open standards, privacy and*

*safety need to be built into the Metaverse from day one" … "you really want to emphasize these principles from the start."*

However, CCDH research on the Metaverse found that VR Chat — the most reviewed social app in Facebook's VR Metaverse — is rife with abuse, harassment, racism and pornographic content. Meta is still applying the "move fast and break things" maxim - to our detriment. Our researchers found that users, including minors, were exposed to abusive behavior every seven minutes. This included:

- Minors being exposed to graphic sexual content.
- Bullying, sexual harassment and abuse of other users, including minors.
- Minors being groomed to repeat racist slurs and extremist talking points.
- Threats of violence and content mocking the 9/11 terror attacks.

We reported all of these identified incidents to Meta using their web reporting tool. All of our reports about users who abused and harassed other users went unanswered.

In the absence of legislation, and in the US, in the face of the section 230 liability shield, the only real incentives to design safe products and services are linked to voluntary actions taken by companies that are motivated by profit. This has led to the current situation where some companies will sometimes take some responsive actions to discrete problems or pieces of content as a result of media, civil society, advertiser, shareholder, or government pressure, but not in a comprehensive, resourced or sustainable way.

Legislation needs to incentivize real safety by design in combination with the other elements of our STAR framework.

## Useful Resources

- CCDH Metaverse research
- The Australian e-Safety Commissioner
- World Economic Forum
- We Protect Global Alliance

## Transparency: algorithms, rules enforcement and economics Introduction

There is currently an information asymmetry and imbalance of power, where technology companies both:

- hold most of the information of what is happening on their platforms and through their services (and how it impacts individuals, communities and society); and
- have an inherent conflict with responding to that harm when the production and spread of that content is integrated into their business model and/or will cost money to address. The business model encourages the spread and amplification of harmful content because profits are made through data collection and high engagement. Harmful content, that is, hate and disinformation, is high-engagement content.

The problem with information asymmetry and the inherent conflict is articulated by Facebook Whistleblower, Frances Haugen, in her testimony to the Senate Committee on Commerce, Science and Transportation:

> "The core of the issue is that no one can understand Facebook's destructive choices better than Facebook, because only Facebook gets to look under the hood. A critical starting point for effective regulation is transparency: full access to data for research not directed by Facebook. On this foundation, we can build sensible rules and standards to address consumer harms, illegal content, data protection, anti-competitive practices, algorithmic systems and more. As long as Facebook is operating in the dark, it is accountable to no one. And it will continue to make choices that go against the common good. Our common good."

An absence of transparency means that truth, reason and democracy is dying in the darkness. Only if we shine sunlight onto what is actually happening through these products and services, can we then anticipate and identify problems and begin to find ways to build system resilience, mitigate and address problems and find solutions. In our report on the Disinformation Dozen, we identified that twelve people were responsible for up to 65% of the anti-vaxx disinformation content spreading on social media - in the heart of the pandemic. The CEOs of Big Tech appeared before the US Congress and promised to take action on those twelve leading super-spreaders of vaccine disinformation. However, when we tracked this a month after these promises were made, we found that they had failed to act on

vaccine disinformation spread by the Dozen that had been viewed up to 29 million times.

The inconvenient truth for Big Tech is that this online hate and disinformation content - like that content identified in our Disinformation Dozen report, while harmful, is profitable and it is only with public transparency and effective research and campaigning from independent organizations like CCDH that these problems are coming to light. The incentives are not in place for proactive, public and full transparency of problem areas and emerging trends - like the emergence of leading disinformation super spreaders and harmful content through COVID-19. Obviously, independent monitoring and evaluation of these transparency efforts is also required (discussed further below under the other STAR elements).

**There are three key areas where transparency should be prioritized:**

**(1) Algorithms;**

**(2) Rules Enforcement; and**

**(3) Economics, particularly related to advertising.**

Transparency will be enhanced where there is a consistent and easily accessible reporting framework across platforms and reporting periods.

## Algorithmic Transparency

### What does this involve?

At a minimum, algorithmic transparency should include:

- **Search algorithms and data** – such as autocompleting a keyword and metadata used;
- **Recommendation algorithms and data** – which curate content that a user may be interested in;
- **Ad-tech algorithms and data** – that target users based on demographics and behavior to optimize advertising; and
- **Moderation algorithms and data** – that target content, users and groups that breach the law or the platform's / search engine's terms and

conditions/community standards. This should include internal metrics, such as the violative view rate.

To help assess the impact of algorithms and products, and to identify emerging forms and trends of harm on platforms, the data above should be supported by public transparency on the most popular content on that platform (with the impact of algorithms controlled and shown). For example, Facebook's top 10 content:

- Most liked.
- Most viewed.
- Most recommended.

Transparency should include publicly accessible data, complemented by more access via a public API, which can be converted into a broader range of formats. There should be clarity about what meta-data is entered into the API to yield particular results. A live public service has the benefits of being faster, giving broader access, providing a public record, and being harder to falsify or mislead.

Within a legislative framework, regulators and courts should have the right to access additional data to ensure legal duties are being complied with.

Individuals should also have a clear right to access and share their own data.

## Why is it needed?

Algorithmic transparency and independent research is critical for understanding what is happening on social media platforms, because these features (and even small changes to these features) drive behavior and outcomes for users. In 2021, we published a report called Malgorithm, which studied the way Instagram's new algorithm[1] recommended content to different users.

We simulated user experience and reviewed recommendations from Instagram's 'Explore' feature and the new 'Suggested Posts' feature, which publishes content

[1] In August/September 2020, Facebook changed the Instagram algorithm to introduce recommended posts and accounts, and a new explore feature. In the past, if you reached the end of your feed of followed accounts, it would signal that you had caught up with posts. With the change to the front page, it started recommending posts that you should read.

into users' feeds based on an algorithm that seeks to maximize engagement and user time on the platform.

We found that more than half of posts recommended by Instagram contained misinformation related to Covid-19, followed by anti-vaxx, QAnon, antisemitism, and election misinformation. By extrapolating from the user profiles we used for the test accounts, we calculated that millions of users are being recommended dangerous misinformation by the algorithm.

We found that Instagram's algorithm was leading users down rabbit-holes to a warren of extremist content:

- Users who engage with anti-vaxx misinformation are being directed to antisemitic posts and election conspiracy theories, while those who engage with QAnon or far-right content are presented with Covid and vaccine misinformation.
- Instagram promoted posts that actually carried warning labels, meaning Instagram's systems screened the content, saw it contained misinformation, yet continued to amplify it.

We found that the algorithm, whether by design or experience of analyzing trillions of clicks, had realized that conspiracy information (as outlined above) is addictive and keeps people on the platform - and engagement time increases profit. Creating a captive audience is good business because eyeballs equal ad revenues equal more money for them.

The Malgorithm study was only possible because of a public announcement about the changes. Other research we do requires access to data and tools, such as APIs. We know from our research that there are limits on the types of information that are currently publicly available, even with knowledge and tools. This information is held by Big Tech companies who can track and trace what is happening online and the impact of changes to algorithms, processes, or safety features. This information asymmetry has devastating impacts for individuals, communities, and society, particularly where the same vested interests who hold most of the information are making all the decisions about content and users, and profiting from light-touch moderation and under-investment in safety.

Transparency is lacking. The status quo unregulated environment means that some companies are only sharing some data, some of the time with some researchers or groups. Like the Big Tobacco or Big Oil lobby, studies which are funded by Big Tech are inherently compromised and there is evidence to suggest that the data currently provided by tech companies is incomplete - see, for example, Facebook's incomplete data sharing with researchers. Facebook Whistleblower Frances Haugen shared internal studies showing that Facebook is well aware that their algorithms are, for example, delivering harmful body dysmorphia content to young girls but are choosing profit over child safety. This information should not be confined to a board room - there is a legitimate and overriding public interest in transparency, which will help to drive safety by design and reduce harm, particularly when coupled with the other elements of our *STAR Framework*.

Our experience is that access to algorithmic data would also be useful for early identification of emerging issues and enables a rich public conversation about the problems on and with social media and what solutions may be available.

Current access to tools for analyzing social media data, such as Twitter's API, are carefully guarded, which impacts the level of scrutiny and insights that can be shared in the common good, and creates a bias for the selection of issues that are highlighted. Other platforms have limited tools or are limiting access to tools. For example, Facebook and Instagram, both owned by Meta, have restricted Crowd Tangle access and are moving towards stopping external access to it altogether, after Meta purchased it from a third party. This trend towards opaqueness is problematic because it will further restrict what information is able to be shared in the public domain on key issues of public importance. Legislation will correct this market failure and ensure that public safety is put ahead of commercial interests.

## Transparency: Rules Enforcement

### What does it involve?

Platforms and search engines need to have clear, accessible and responsive complaints'/reporting systems, where terms and conditions/policies ("rules") have been breached. Transparency on rules enforcement means providing public access and data on:

- Rules: content of terms and conditions, reporting pathways, and content moderation policies, practises, and tools; and
- Enforcement: on how terms and conditions and community standards have been breached; which rules are applied in terms of prioritization and criteria, and when enforcement action does or does not take place. This data should include both overall violation rates of rules and by particular topics (e.g. COVID vaccine misinformation).

Currently, in most countries transparency reports on content moderation and design choices are provided by technology companies on a voluntary basis.[2] The UK Government noted that these voluntary reports, where they exist:

> "... often provide limited detail across important areas including content policies, content moderation processes, the role of algorithms in moderation and design choices, and the impact of content decisions."

In addition, a common issue that we have experienced through our work is that, while companies may release a transparency report that states the total number of individual pieces of content related to a specific policy that has been removed or otherwise moderated, there is no data provided on what proportion of that type of content it comprised. The extent of this disparity between what was stated and what was known was also evidenced by Whistleblower Frances Haugen, who advised that internal estimates were that Facebook may action as little as 3-5% of hate and about 6/10 of 1% of violence and incitement content on Facebook.

## Why is it needed?

Transparency in rules enforcement is both a consumer protection and a rich source of data for understanding the nature, extent and trends of online harm, and the impact and effectiveness of any changes or interventions. For public safety, users need to easily understand how they can make a report about harmful content or another user, any grounds for a response, and appeal rights. Transparency is an important component of an effective reporting system to help ensure that harmful content is removed and bad actors are held to account. The added accountability through transparency will also mean that more thought and investment will go into

[2] The Digital Services Act and the UK's Online Safety Bill when operational will change this situation for countries in scope of those laws.

the design and responsiveness of these systems, particularly when coupled with the other elements of our *STAR Framework*.

Importantly, policies have no weight if there is no enforcement. For example, in our study Digital Hate: Social media's Role in Amplifying Dangerous Lies about LGBTQ+ People, we note that Twitter and Facebook both have existing rules against hate speech - including the term "groomer" - but enforcement is lacking (also see Daily Dot). We identified 989,547 tweets posted between January 1st and July 27th that mention the LGBTQ+ community alongside slurs such as "groomer", "predator" and "pedophile". Our audit found that Twitter failed to act on 99% of the 100 hateful tweets reported to them anonymously by CCDH researchers. Meanwhile, Meta is profiting from ads promoting rhetoric that the LGBTQ+ community and its allies are 'grooming' children. CCDH identified at least 59 ads promoting this rhetoric, which have been served to users over 2.1 million times.

This is not a one-off problem, our research shows that the self-regulatory environment is leading to these companies failing to act on reported harmful content at least 85% of the time. This is harmful content that is actually brought to their attention as breaching their policies using their own reporting systems, let alone content that they should be proactively finding and addressing. For example, in our study on Anti-Muslim Hate: Failure to Protect, we found that platforms collectively fail to act on:

- 89% of Islamaphobic and anti-Muslim hate content on their platform;
- 97% of Islamophobic posts — even after they are flagged to moderators; and
- 89% of posts promoting the "Great Replacement" conspiracy theory—violating pledges made following the 2019 Christchurch mosque terror attacks and signing on to the Christchurch Call.

Not only are Big Tech failing to act on reported content, they have also failed to design ways to report all forms of content or abusive users. In our report Hidden Hate, we found that Instagram, owned by Meta, failed to act on 9 in 10 reports of misogyny and violent threats over direct messages reported using its tools, and failed to act on any image-based sexual abuse within 48 hours. It was not possible to report voice messages. Similar issues of inaction and system failure for reporting abuse arose through the course of our Metaverse study. While the VRChat app marks users who are logged in using an Oculus Quest in their "nameplate", making it

possible to report Oculus users to Meta, some users marked as Quest users appeared to have usernames that did not match any existing Oculus user profile. In those cases, reporting systems in the Oculus operating system and on the Oculus website refused to accept a report. This is incredibly problematic given our finding that there was one incident of abuse and harassment every seven minutes and this product was marketed as being safety first and family-friendly.

Through our work, we have experienced additional issues with reporting accounts and hashtags. There are also issues where the subject of a complaint does not fit into the platform's predefined categories of grounds for reporting. Though options and avenues for reporting have improved over time, there remain differences in reporting quality. There are no common standards or expectations.

Other regions and countries, like the EU and the UK, are developing ways for the independent regulator to monitor the effectiveness of complaints and reporting systems. Transparency and information outlined above will be critical for this accountability function. In the UK and Australia, this information will also be relevant for the development of new and revised risk assessments and codes of practice.

The design and use of rules on platforms and search engines are also problematic. With existing rules, there's generally a reliance on extremely broad standards and then after media attention, there may be a claim that the existing rules apply to additional issues. For example, Facebook, Twitter and Google all have standards of harmful health information that they applied to COVID disinformation after campaigning, research, and pressure from the public, civil society, and officials. This was a delayed reaction to a known and potentially deadly problem.

## Economic Transparency

### What does this involve?

Legislation that requires greater transparency over ads: specifically, understanding where, when, by whom, and using which data.

One option for achieving this is to require advertisers to publicly declare, on their websites, the domains where their ads appear. This creates a driver for corporate accountability, i.e. that consumers' money is not being funneled to content that

fundamentally harms individuals, communities and society. This type of information is often provided to advertisers by brokers, some of which are updated in real time.

This requirement would simply ensure that advertisers disclose the URLs of the pages on which their adverts appear, but not other information, such as performance data or targeting criteria. It would not create a duty for advertising organizations to conduct costly studies—by making these URLs publicly available, it will make it easier for researchers, journalists, authorities and the public to instantly access the relevant information. This creates an accountability ecosystem of enabling legislation, transparent corporate behavior, and civil society and other companies doing the checking. There are organizations such as the Global Disinformation Index and NewsGuard that can provide the "checklist" for advertisers. CCDH's Stop Funding Misinformation has a much shorter and much more focused "Blacklist".

## Why is it needed?

Google, Twitter and Facebook all have ad libraries. All have limitations. For example, Google only provides a record from the past 30 days; Facebook only lists political and social issue ads (which may lead to many not being included, depending on how the definition is interpreted). Google and Twitter both provide poor transparency in this area. Tiktok does not have an ad library and provides no transparency for ads on its platform. The status quo allows Big Tech to profit from both hate and the controversy surrounding that hate, which drives attention and traffic to their platforms and makes money from ad revenue.

Each year, respectable companies and their customers unwittingly funnel millions of pounds directly to the Internet's most malicious and subversive actors and messages. Misinformation and hate sites are almost entirely funded by online advertising — often paid for by unsuspecting mainstream organizations who don't know what content their brand is appearing next to, and thereby funding. Their adverts are placed by third-party Brokers, such as Google's Adsense business, which then allocate adverts to particular sites to fulfill predetermined target demographic (age/gender/location) and psychographic (attitudinal and behavioral) profiles. The use of algorithms to select which ads appear where to fulfill a target profile has led to these services being called "programmatic advertising."

In their zeal to maximize the real estate that they can place ads on, however, brokers often agree to place ads on content even if some of it is dangerous or offensive. As misinformation and hate sites are one step removed from their advertisers, this advantages websites that prioritize "clickbait"—stories that generate the most traffic, irrespective of the veracity of the information presented, or even if they peddle hate and division.

The presence of mainstream, respectable organizations next to extremist content—including climate denial, anti-vaxx propaganda, political misinformation, incel ideology and racial hatred—also serves to normalize extremism. CCDH found that adverts for brands such as *Chevrolet, Capital One, DHL International, Boots, Canon, Sensodyne, Paperchase, Bloomberg, the International Rescue Committee* have been automatically placed on sites dedicated to promoting hate and/or misinformation.

It also legitimizes harmful disinformation and has led to dangerous offline behavior. For example, we know that websites promoting the following disinformation have profited from these ads:

- **Anti-vaccine and Covid-skeptic misinformation**
- **Climate denial** and conspiracy theories about climate activism;
- **Conspiracy theories promoting the "stolen election" myth** in the wake of the 2020 US presidential election, which directly led to the domestic terror attack on the Capitol on January 6 2021; and
- **Misinformation sites** promoting antisemitism, Islamophobia, misogyny and pro-Assad conspiracy theories.

More information about this problem is available here: VIDEO: Stop Funding Misinformation / www.counterhate.com.

## Accountability to democratic and independent bodies

### What is it?

Regulation is most effective where there are accountability systems in place, particularly where there is a risk of inaction because of profit motives and commercial factors. Accountability systems usually include enforcement, which may include a range of regulatory tools and interventions, and/or an independent pathway for challenging decisions or omissions made by technology companies. A comprehensive system would have a body that can deal with systems and underlying drivers of harm, with a broader range of intervention tools. We envision this body having key relationships with civil society, communities and researchers and functioning internationally with partners as part of the global effort to address online harm and misinformation.

The choice of body may vary between jurisdictions and the legal culture of a country. In many jurisdictions, for example, there is an independent regulator (e.g. Ofcom in the UK) with further appeal rights to independent judicial bodies/courts, whereas in the US the standard regulatory approach is a regulator body headed by political appointments with the ability to appeal decisions to tribunals/judges or a direct pathway to litigate actions through the courts. The key essence, however, is that there is a body independent and separated from the technology companies themselves who are ensuring that obligations, whether through contract and negligence law or statutory duties, are being upheld.

### Why is it needed?

Self-regulation has clearly failed. Given the record of facilitating and profiting from harm, technology companies cannot be trusted to oversee and enforce their own rules. An accountability system provides appropriate checks and balances on corporate power and the perverse incentives that currently exist for taking a light-touch approach to safety by design, transparency and rules enforcement.

# Responsibility for platforms and their senior executives

## What is it?

The final element of the *STAR Framework* is responsibility. Both social media platforms and search engine companies and their senior executives that are (and should be) responsible for implementing duties under a legislative framework or otherwise being held to account through the courts. A dual approach - targeting both companies and their senior executives - is a common intervention strategy for changing corporate behavior - see, for example, health and safety and corporate manslaughter law.

When a platform fails to act to remove harmful content, even after warnings from the regulator, discharge transparency or other requirements under the legislation, and/or fails to deploy its vast resources to avoid harms generated on their platform, courts or a regulator could be free to decide if someone is harmed by their inaction. This creates an economic incentive for action. In the US context, the risk of litigation, let alone litigation itself, is a strong motivator for shifting corporate behavior.

Senior executives need to be held accountable for their conduct as administrators of platforms that hold enormous power over discourse not just in terms of content moderation, but also the amplification of content, institutional and user experience design of the systems through which discourse occurs, and inequity in user experience for marginalized communities.

Offenses supported by sufficient financial penalties are often a good motivator. For example, in Germany, under the "NetzDG law", platforms which fail to swiftly remove harmful content face significant financial penalties. Opponents warned that the NetzDG law would lead to overly sensitive censorship and infringements on free speech. So far, it has not. Once there is a financial incentive to comply with local laws, the social media companies are suddenly able to act: it is no coincidence that Facebook was quick to establish a vast hub of moderators in Berlin. The UK is introducing a similar penalty system in its Online Safety Bill, which would penalize

companies up to 10% of their gross global turnover and has separate offenses for senior managers in social media companies.

## Why is it needed?

The financial, labor, and emotional burden should not fall on individuals to address online harm, particularly where they don't have the privilege of transparency and access to the underlying causes of that harm or the technical skills, access or resources to design interventions that would change the operating environment. But this is the status quo.

Corporate responsibility that is clear and enforced with the right incentives and disincentives in place will change corporate behavior. There should be corporate responsibility for products and services that are dangerous and facilitate harm and for failure to act on clear duties of care and compliance. Increasing the costs for providing unsafe products and services will mean that the companies and senior management within those companies start to make different decisions. There must be consequences for wrongdoing and harm.

Social media should be made safe before being available to users, including children, in the same way as we demand from those producing food, cars or pharmaceuticals. Removing any general and unjustified exceptions to civil law to drive a rebalancing of the investment that is spent on engagement with safety features will lead to a safer environment for all users, including children.

## Next Steps

The STAR Framework is intended to inform stakeholders and governments advocating for and designing legislative reform. Global standards help to ensure the efficiency, effectiveness and impact of national efforts, and are best supported by a strong relationship with independent civil society and researchers.

For more information, requests for briefing, or any other queries should be directed to info@counterhate.com.