Researching the Evolving Online Ecosystem: Barriers, Methods and Future Challenges

Jakob Guhl, Oliver Marsh & Henry Tuck

About this publication

This report outlines the findings from the initial scoping phase of a project supported by a grant from Omidyar Network and launched by the Institute for Strategic Dialogue (ISD) and CASM Technology to identify online spaces used by extremist, hate and disinformation actors and communities as they increasingly move away from mainstream social media platforms. The report outlines the key barriers posed by these platforms to researching and mitigating harmful content and behaviours, and reviews existing research methodologies and tools to address these barriers. Finally, the report presents possible future scenarios for the evolving online ecosystem, and proposes a series of initial recommendations for policy-makers, platforms and the research community.

Acknowledgments:

This report would not have been possible without funding support from Omidyar Network. We would like to express our gratitude to Wafa Ben-Hassine, Anamitra Deb and Emma Leiken for their vision, continuing support and insightful feedback.

The authors would also like thank the wider project team for their contributions that have made this report possible: Francesca Visser, Jacob Davey, Lea Gerster, Daniel Maki, David Leenstra and Francesca Arcostanzo at ISD, and Nestor Prieto Chavana and Carl Miller at CASM.

Finally, we would also like to thank Eduardo Ustaran and Nick Westbrook at Hogan Lovells for their invaluable time and support in understanding the legal challenges addressed in the report.

About the authors

Jakob Guhl is a Senior Research Manager at ISD, where he works within the Digital Research Unit and with ISD Germany. His research focuses on the far-right, Islamist extremism, hate speech, disinformation and conspiracy theories. Jakob has been invited to present his research to the German Ministry of the Justice and provided evidence to the German Minister of the Interior on how to strengthen prevention against right-wing extremism and antisemitism.

Oliver Marsh is the founder of The Data Skills Consultancy, which supports work at the intersection of data skills and soft skills. Previously as a government official, he helped create the Rapid Response Unit in Downing Street and the UK's post-Brexit Data Adequacy capability in DCMS. He is a Fellow of the think-tank Demos, a Policy Fellow of the Royal Academy of Engineering, and an Honorary Research Associate of the Science and Technology Studies Department at UCL.

Henry Tuck is the Head of Digital Policy at ISD, where he leads work on digital regulation and tech company responses to terrorism, extremism, hate and dis/misinformation online. Henry oversees ISD's Digital Policy Lab (DPL) programme and advisory work on key digital regulation proposals in Europe and Five Eyes countries, and collaborates with ISD's Digital Analysis Unit to translate research into actionable digital policy recommendations.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Glossary

Alt-tech describes social media platforms used by groups and individuals who believe major social media platforms have become inhospitable to them because of their political views. This includes platforms built to advance specific political purposes; libertarian platforms that tolerate a wide range of political positions, including hateful and extremist ones; and platforms which were built for entirely different, non-political purposes like gaming.

An **Application Programming Interface (API)** is a software intermediary that allows two applications to communicate with each other. They have a huge range of uses; however, in the context of this report, APIs allow researchers to access certain data from some online platforms via requests. As an intermediary, APIs also provide an additional layer of security by not allowing direct access to data, alongside logging, managing and controlling the volume and frequency of requests.

Conspiracy theories are attempts to explain a phenomenon by invoking a sinister plot orchestrated by powerful actors. Conspiracies are painted as secret or esoteric, with adherents to a theory seeing themselves as the initiated few who have access to hidden knowledge. Supporters of conspiracy theories usually see themselves as in direct opposition to the powers who are orchestrating the plot which are typically governments or figures of authority.

ISD defines **disinformation** as false or misleading content that is spread with the intent to deceive, or secure economic and/or political gain, and which may cause public harm. When referring to such content that is spread unintentionally, we will be using the term **misinformation**.

Encryption refers to the process of encoding information so that it is rendered incomprehensible to anyone except specified receivers.

ISD defines **extremism** as the advocacy of a system of belief that claims the superiority and dominance of one identity-based "in-group" over all "out-groups." It advances a dehumanising, "othering" mindset incompatible with pluralism and universal human rights.

We define **fragmented platforms** as those where online content is theoretically accessible, with no technological or ethical barriers, but nevertheless cannot be searched quickly or systematically, for example, via an API. Relevant content must therefore be found manually among vast amounts of other material.

We use **harmful content and behaviours** to refer to a broad spectrum of online activities that can have a negative impact on human rights, society and/or democracy. These can range from targeted harassment of individuals, to incitement of violence against a particular group, or the spreading disinformation and harmful conspiracy theories. In some instances, the risk of harm may be intrinsic to the content itself, with the risks exacerbated by amplification. In other instances, the risk of harm may be caused by aggregate patterns of behaviour rather than the nature of the content itself. Depending on the geographic and legal context, different forms of harmful content or behaviours may or may not be illegal. Depending on the platform, harmful content or behaviours also may or may not be covered by a company's 'Community Guidelines', standards or rules.

Hate is understood to relate to beliefs or practices that attack, malign, delegitimise or exclude an entire class of people based on protected characteristics, including ethnicity, religion, gender, sexual orientation or disability. Hate actors are understood to be individuals, groups or communities which actively and overtly engage in the above activity, as well as those who implicitly attack classes of people through, for example, the use of conspiracy theories and disinformation. Hateful activity is understood to be antithetical to pluralism and the universal application of human rights.

Open platforms are social media platforms on which content is visible to general users without further verification and often accessible via search engines. By contrast, content on **closed platforms** will not be easily accessible via search engines and often requires additional authentication or an invitation. Platforms will often contain both open and closed elements, for example, Facebook has public (open) and private (closed) groups.

Introduction

Many extremist, hate and disinformation actors and communities are moving away from mainstream social media platforms. Instead, they are adopting a wider and more diverse range of online spaces that offer even less moderation, or exploiting platforms that offer greater privacy, security or anonymity. This report outlines findings from the initial scoping phase of a project that was funded by Omidyar Network and launched by the Institute for Strategic Dialogue (ISD) and CASM Technology in order to identify these online spaces and establish research methodologies to monitor and analyse them.

Phase II of the project will apply the findings from the scoping phase to the research of three small platforms in English, French and German in order to expand the research field's understanding of which methodologies (with existing data access) are applicable to these online spaces. During the third and final phase of the project, ISD will share the lessons of Phases I and II with policy constituencies and host an expert roundtable to share the research findings and implications for platform transparency and data access with relevant government, regulatory, research and private sector representatives. Based on our findings, we will also consider how the legal and regulatory landscape may need to adapt to keep pace with the increasing range and technological variety of online platforms, while also respecting and protecting vital rights to privacy, security and anonymity online.

Our ultimate aim is to understand and counter the spread of harmful content and behaviours online. Spreading harm through mediums of communication has always taken many forms, from plans made via private letters to agitation in public squares. However, recent decades have seen an important technological revolution: the increasing ability to systematically collect, store and precisely search communications data. Initially, this required specialised access to data and so was largely limited to select groups (e.g. owners of communication technologies or intelligence agencies). The rising popularity of public online spaces, particularly a handful of dominant social media platforms, has allowed a wide range of researchers to track, analyse and, hopefully, counter various forms of online harms. But this trend may now be reversing. Multiple social and technological shifts – the growth of platforms ideologically opposed to moderation; the emergence of new technologies (e.g. blockchain, augmented and virtual reality (AR/VR), and artificial intelligence); and the increasing adoption of encrypted platforms for private messaging – may be combining in ways that make harmful online activity harder to address.

This report considers these challenges as well as the methods and tools available for researchers to address them. After a general introduction to the task of identifying harmful content and behaviours online, Section 1 outlines the process and findings of our scoping exercise aimed at mapping the current landscape of online platforms and apps popular among harmful communities. Based on this exercise, Section 2 introduces three types of barriers to research or data access posed by these platforms; it will also consider the current and (potential) future implications of these barriers for the research community, policy-makers and companies. In Section 3, we summarise three broad types of research methodologies for finding harmful content and behaviours online; the potential strengths and weaknesses of each methodology in tackling each barrier; and the tools available for researchers to investigate harmful content and behaviours on smaller platforms. In Section 4, we propose case study platforms and potential research approaches for overcoming these barriers, which will be trialled during Phase II of this project. In Section 5, we present possible future scenarios for researchers and those combating harmful content and behaviours online, from pessimistic to optimistic, and propose a series of initial recommendations for policy-makers, platforms and the research community. Finally, the accompanying annexes to this report provide the full results of the platform-scoping exercise and further explore possible ethical, legal and security risks associated with researching these online platforms.

Harmful Content and Behaviours Online

Harmful content and behaviours can span a wide spectrum of activity, from online harassment and the incitement of violence to the spreading of disinformation and harmful conspiracy theories. The risk of harm may be intrinsic to pieces of content themselves; in other instances, the harm may be caused by patterns of behaviour rather than the nature of the content itself. In the case of harmful behaviours online, individual items of content may not be particularly harmful in isolation, but

the systematic amplification of unverified information or polarising narratives may prove harmful in the aggregate. Harassment is a particularly pertinent example of this; while one-off pieces of highly provocative or antagonistic content may not result in significant harm, if they are part of a pattern of behaviour that targets particular individuals or communities in large volumes or over an extended period of time, this harassment can deter journalists, activists, politicians or members of marginalised communities from participating in the online public sphere.

Such content and behaviours may violate the human rights of targeted marginalised communities, undermine trust in democratic institutions and principles, and make it very difficult to find common ground in political debates. It can encompass overtly ideological content (e.g. violent extremist content), broader societal issues with political implications (e.g. misogynist "incel"[i] content) or even non-political but harmful content (e.g. promoting self-harm). Depending on the geographic and legal context, different forms of harmful content and behaviours may or may not be illegal. While some forms of harmful content are illegal in most contexts (e.g. terrorist content or child sexual abuse material), laws around other forms of harmful content, such as hate speech, can vary considerably across national jurisdictions. Many actors spreading harmful content are also aware of legal boundaries and are careful to use coded or implicit language to avoid crossing into illegality. The growing recognition that many forms of legal content can still result in significant harm has led to discussions around how to address harms like dis/misinformation through regulation, such as the EU's Digital Services Act[1] or the UK's Online Safety Bill.[2]

Private sector companies also set their own 'Community Guidelines', standards or rules that outline the types of content and behaviours that are allowed on their platforms. At a minimum, these terms or guidelines typically cover illegal content or behaviours in the jurisdictions in which the companies operate. Moreover, many major social media platforms also choose to go further, prohibiting other forms of harmful but legal activity. While their precise definitions, thresholds and enforcement approaches may differ, many of the largest companies' guidelines, standards or rules have converged to prohibit a similar range of legal but potentially harmful activity under pressure from advertisers, civil society, legislators and users.[3]

In contrast, we have identified in our research considerable diversity in the community guidelines, standards or rules of many of the smaller platforms that make up the broader online ecosystem. Different platforms can take radically different positions on various forms of so-called "legal but harmful" activity. Some may only prohibit illegal activity in the jurisdiction in which they are based, while others may choose to go further. This variance can be due to several factors. Some platforms may lack sufficient resources to implement and enforce more comprehensive rules (e.g. platforms that make little or no revenue or profit). Other platforms may have more fundamentalist commitments to absolute freedom of speech or may believe such a stance will attract a certain type of user. Additionally, there are also some platforms that adopt a more ideological position, for example, those purpose-built to cater to extremist communities (e.g. far-right extremist forums, such as Iron March or Fascist Forge).[4]

If harmful content and behaviours are identified quickly enough, it may be possible to limit the harm they can cause through legal, technical or other measures. For example, platforms can take a range of measures to remove or restrict the relevant content or accounts where their rules are broken[5], or the creator could face charges under domestic legislation if they are deemed to have crossed into illegality. Even if the content is already circulating, finding harmful content can help develop effective counternarratives and, hopefully, slow the spread. More broadly, being aware of harmful activity can highlight trends, techniques and tools used to develop such messages, and therefore help to more effectively predict, spot and counter harmful content and behaviours in general.

i Incels (short for "involuntary celibate") are an online subculture whose predominantly male adherents believe themselves to be unable or too undesirable to enter into sexual relationships. Incel communities often propagate highly misogynist ideas, and adherents of the subculture have engaged in mass-casualty attacks; see O'Donnell, Catharina and Shor, Eran, '"This is a political movement, friend": Why "incels" support violence', The British Journal of Sociology, 73(2), January 2022, https://onlinelibrary.wiley.com/doi/10.1111/1468-4446.12923.

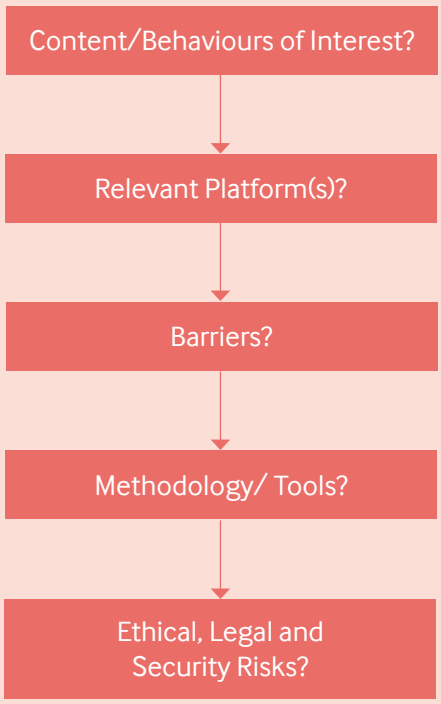
Finding Harmful Content and Behaviours

Digital technologies have greatly contributed to the ease of locating and collecting data. Searching has been made vastly more powerful by a range of tools: search engines like Google; platform-specific technologies like CrowdTangle[ii] or Twitter Advanced Search; marketing-focused social media listening tools like Brandwatch; and research-focused technologies like Method52.[iii] Monitoring a wide range of online spaces is also much easier than in-person infiltration of numerous extremist groups.

Multiple research approaches can support one another. For instance, searching for specific keywords or content may lead researchers to a new online space where they discover new keywords or topics to search for, and so on. This is particularly important for finding and addressing harmful content online. Such content often develops in specialised spaces (e.g. extremist forums) before being pushed onto mainstream platforms where it can acquire much greater spread and new audiences. Similarly, it has been documented that harassment campaigns are often coordinated in such specialised forums.[6] A combination of observing niche online spaces and searching for content as it spreads is therefore important to following (and hopefully shortening) the life-cycle of harmful content; however, barriers to either discovering specific content and behaviours or accessing broader online spaces can break this virtuous cycle.

Barriers to finding content and identifying harmful behaviours online may be technological, social and/or legal in nature. Online platforms may be deliberately designed to minimise access to data, or this may be a side-effect of other features such as end-to-end encryption. It should be emphasised that such features aimed at protected, private and secure communication have major upsides from a human rights and privacy rights perspective. In authoritarian countries (though not just there), secure communication technologies protect activists and dissidents from surveillance and government infringement. Combatting harmful activity on platforms using such technologies should not come at the price of sacrificing these benefits.

Figure 1: Key research decisions for studying harmful online communities, content or behaviour



When embarking on a project aimed at studying harmful communities, content or behaviour online, a series of key decisions need to be made. First, which harmful actors, communities, behaviours, dynamics or narratives are of interest? Second, on which platform(s) do we expect to find them? Third, which barriers to research do the platform(s) of interest present? Fourth, what methodologies and tools are available to overcome these barriers? And finally, what ethical, legal and security risks may arise from these decisions?[iv] While these issues and decisions are discussed in separate sections of this report, it should be noted that these processes will often not be linear but instead run in parallel, directly affecting each other.

There is an argument that, for a variety of reasons, barriers to researching harmful content and behaviours online are increasing. This problem appears to be particularly urgent in those online spaces that offer less moderation

ii A Meta-owned tool that provides access to some (increasingly limited) publicly available data from Facebook and Instagram.

iii Method52 is a social media analysis tool developed by CASM and the University of Sussex. For more information see 'Technology and Values', CASM, https://www.casmtechnology.com/pages/technology.

iv See the annexes to this report for detailed discussions of such risks for researchers and organisations.

and/or greater privacy, security or anonymity. To gain an overview of the current landscape of platforms and apps popular among harmful communities, we developed a list of case study platforms from three recent English, French and German datasets focused on extremism or harmful conspiracy theories. The process for identifying these platforms as well as the findings are outlined in Section 1 below. Based on the platforms identified during this platform-scoping phase, Section 2 of this report presents three broad types of barriers to researching and combating harmful content and behaviours on these platforms.

Section 1: Platform Scoping

Regarding the issues identified in the introduction, ISD initially compiled a list of platforms and apps referred to by different harmful communities in 2021 in order to identify any new and emerging platforms. Furthermore, barriers for finding harmful content on these platforms were recorded and categorised.

In order to conduct this analysis, ISD used a seed list of actors and communities on Facebook, Instagram, Twitter, YouTube, Reddit, 4chan, Telegram and Gab. This list was gathered from previous research projects on disinformation, hate and extremist groups in French,[7] English[8] and German.[9] These datasets, compiled in 2021, included lists of actors and groups that were found to have spread disinformation and conspiracy theories about COVID-19 and vaccines, and/or to have participated in far-right extremist or antisemitic activities.[v] Using these datasets, ISD was able to identify any links to other platforms shared in these groups. This exercise allowed us to list systematically the most common platforms to which the communities were linking.

This methodology has some caveats. The starting point of this exercise only included platforms and communities that were already accessible to researchers, and it did not comprise closed and encrypted platforms or closed messaging apps. Moreover, there was a focus on far-right and conspiracy actors in our seed list; different communities and groups (e.g. Islamist extremists) may also be migrating from mainstream platforms but towards a different range of alt-tech platforms. Finally, our seed list was focused on English, French and German language online communities. There are likely to be other emerging platforms that are relevant in other languages and country contexts. For these reasons, the results cannot be representative of the entire online disinformation and hate landscape but are restricted to the communities and the languages included in the analysis.

An alternative approach could include a wider variety of platforms as a starting point, for example, closed or encrypted messaging platforms like WhatsApp. This approach, however, would give rise to additional ethical and legal concerns. Users utilise these services under the assumption that their conversations are private. Gaining access to these closed spaces might present ethical (and potentially legal) risks due to the additional levels of deception and/or intrusion that may be required. Alternatively, a selection of platforms could have been derived from existing literature and ethnographic research into the identification of alt-tech platforms prone to exploitation by extremist groups (though this risks bias towards the most high-profile platforms). Both approaches could be used in future to supplement the list we have compiled; however, for the purpose of making initial investigations into the broad threats we have identified, this list was more than sufficient.

The collection resulted in 35 platforms in French-speaking countries, 31 in German-speaking countries and 21 in English-speaking countries.[vi] In order to identify different types of barriers to research, these platforms were categorised based on their content, technological features, scope and relevant policies. We also considered the platforms' attitudes towards privacy and free speech, assessed via their creators' expressed views, the companies' policies and/or the nature of their user base as key elements in the categorisation.

In order to narrow down this initial list of platforms and identify the most relevant for our research, we developed a coding sheet and coded each platform for its features. The coding sheet included general information on each platform, such as the number of users globally, the purpose of the platform, when it was founded and whether it presents clear content policies, particularly around hate speech and disinformation. We also identified whether each platform has terms and conditions regarding data usage by external parties and whether the platform offers closed groups.

Technological features of each platform were identified to assess any barriers for conducting analysis. These features include whether the platform has a search function and/or an API, and whether it is encrypted or makes use of new technologies like AR/VR or blockchain. Finally, barriers for finding harmful content were noted and categorised into three types (expanded on further in the following sections):

v As the datasets were drawn from recent but distinct projects, the date range and sizes were varied. The English data included 2.5 million posts between 1 January 2021 and 30 November 2021. The German data included 659,000 posts between 1 January 2021 and 12 September 2021. The French data included 2 million posts between 31 July 2020 and 31 January 2021.

vi See Annex: Platform-Scoping Data – Link Counts for the full list by language.

- Technological features which block/limit access to data
- Ethical and legal issues faced by researchers
- Fragmentation of content across platform(s) in a way which impedes efficient and systematic data collection

As the aim of this exercise was ultimately to identify barriers to research, we restricted our selection of platforms to the ones that presented at least one of the three barriers. This resulted in 15 platforms in total across the three languages. Among these platforms, we included:

- Traditional social media and messaging apps with closed groups like Facebook, VK, Telegram and WhatsApp as the presence of private groups gives rise to additional ethical challenges.
- Discord as it presents ethical barriers (in its closed groups) and fragmentation barriers (in its public groups because research on the platform can only be done server by server and not in a systematic way).
- Odysee as it presents both a fragmentation and a technological barrier.
- Kik as the content of chats is not accessible with existing methods and tools, presenting a technological barrier.
- A range of other platforms that have both a technological and an ethical barrier (nandbox, Hoop Messenger, Riot, Minds and Rocket.Chat).
- Vimeo, DLive, and Spotify as limitations in analysing audio-visual content (and in the case of DLive, the use of blockchain technology) present technological barriers.

English-language

	Telegram	Minds	Discord	Facebook	VK
Leadership	Pavel Durov (CEO)	Bill Ottman (CEO)	Jason Citron (CEO)	Mark Zuckerberg (CEO)	Vladimir Kiriyenko (CEO)
Number of global users	500 million	2.5 million	350 million	2.89 billion	460 million
Clear content policies?	Policies only against promoting violence and illegal pornography	Yes	Yes	Yes	Policies against terror, propaganda and hate speech but not disinformation
Purpose	Alternative chat platform to avoid government surveillance	Alternative to Facebook, which mines a large amount of data	Communication for gamers	Social networking	Social networking
Year of Founding	2013	2011 (launch in 2015)	2015	2004	2006
Terms and conditions for data usage?	Yes	Prohibits data export	Does not allow data mining or extraction	Yes	Yes
Embedded analytics?	No	Yes	Yes	Yes	Yes
Domain registration record available?	Yes	Yes	Yes	Yes	Yes
Closed groups?	Yes (e2e-encryption on chats)	No	Yes	Yes	Yes
Fragmentation barrier?	No	No	Yes	No	No
Ethical and legal barrier?	Yes, closed groups	Yes, closed groups	Yes, closed groups	Yes, closed groups	Yes, closed groups
Technological barrier?	No	Yes, e2e-encryption and blockchain	No	No	No
Search box?	Yes	Yes	Yes	Yes	Yes
API?	Yes	Yes	Yes	Yes	Yes
Link to API	https://core.telegram.org/	https://gitlab.com/minds/engine	https://support.discord.com/hc/en-us/articles/212889058-Discord-s-Official-APIcles/212889058-Discord-s-Official-API	https://developers.facebook.com/docs/pages/	https://vk.com/dev
Encrypted?	Groups and channels use cloud encryption; chats use e2e-encryption	Yes	Yes. Standard encryption	No	No
New technologies?	No	Yes, blockchain	No	Yes, VR	No
Notes		Collects statistics on user behaviour. Mines data of most popular accounts and occasionally makes it public. Does not disclose personal information.			

German-language

	DLive	Hoop Messenger	nandbox	Odysee	Riot/Element	Rocket.Chat	WhatsApp
Leadership	Justin Sun (CEO)	Sahand Adilipour (president)	Hazem A. Maguid (CEO)	Julian Chandra (CEO)	Matthew Hodgson (CEO and CTO)	Gabriel Engel (CEO)	Mark Zuckerberg (CEO)
Number of global users	5 million	Unclear	Unclear	8.7 million	35 million	12 million	2 billion
Clear content policies?	Yes	Yes	Yes	Yes, but not for disinformation	No, but offers guidelines for moderators	Yes, but not for disinformation and hate speech	Yes
Purpose	Live streaming	Secure messaging	Secure messaging	Decentralised video-sharing platform	Decentralised, secure messaging	Secure messaging	Messaging
Year of Founding	2017	2014	2016	2020	2016 (as Riot)	2015	2009
Terms and conditions for data usage?	Yes, shares data with third parties	Does not share data unless legally required	Does not share commercial data but does co-operate with law enforcement	Does not share personally identifiable data but provides anonymised data	Only in exceptional circumstances to comply with the law	No	Shares data with other Meta companies and third parties
Embedded analytics?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Domain registration record available?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Closed groups?	No	Yes	Yes	No	Yes	Yes	Yes
Fragmentation barrier?	No	No	No	Yes	No	No	No
Ethical and legal barrier?	No	Yes	Yes	No	Yes	Yes	Yes
Technological barrier?	Yes, audio-visual	Yes, encryption	Yes, encryption	Yes, audio-visual	Yes, e2e encryption	Yes, encryption	Yes
Search box?	Yes	Yes	Yes	Yes	Yes (public rooms)	No	No
API?	Yes	No	Yes	No	Yes	Nes	No (for the most part)
Link to API	https://docs.dlive.tv/api/	n/a	https://api.nandbox.com/#nandbox-api	n/a	https://element.io/developers	https://developer.rocket.chat/reference/api	https://www.whatsapp.com/business/api
Encrypted?	No	Yes	Yes	No	Yes	Yes	Yes
New technologies?	Yes, blockchain	No	No	Yes, blockchain	Yes, decentralised protocol	No	No
Notes		Channels can be deleted.					

French-language

	Spotify	Vimeo	Kik
Leadership	Daniel Ek (CEO)	Anjali Sud (CEO)	Ted Livingston (CEO)
Number of global users	173 million (premium subscribers)	175 million	300 million
Clear content policies?	Yes, but not for disinformation	Yes, including disinformation on selected topics	Yes
Purpose	Audio streaming	Video hosting and sharing	Messaging
Year of Founding	2006	2004	2010
Terms and conditions for data usage?	Shares anonymised data with researchers	No	No
Embedded analytics?	Yes	Yes	Yes
Domain registration record available?	Yes	Yes	Yes
Closed groups?	No	No	No
Fragmentation barrier?	Yes	No	No
Ethical and legal barrier?	No	No	No
Technological barrier?	Yes, audio material	Yes, audio-visual material	Yes, content not accessible
Search box?	Yes	Yes	Yes, but for users not content
API?	Yes	Yes	Yes
Link to API	https://developer.spotify.com/documentation/web-api/	https://developer.vimeo.com/api/reference	https://kik.readthedocs.io/en/latest/api.html
Encrypted?	Yes, music	No	No
New technologies?	No	No	No

Section 2: Three Barriers

In this section, we present three broad types of barriers to research. These are not mutually exclusive. Although we primarily focus on each type of barriers' impact on finding harmful content and behaviours, each additionally creates challenges for moderating or mitigating the impact of such activity; we will briefly introduce some of these challenges too. There are very few cases where these barriers make conducting research on a given platform entirely impossible. In the next section, we explore a series of methods and tools that can help to overcome these barriers.

Barrier Type 1: Technological

Technology can greatly improve access to data. It can also limit it. Platforms may deliberately use technologies which restrict access to data, or they may also have other technological features which inadvertently create barriers for researchers. The technological features of specific forms of content may also restrict researchers' ability to conduct systematic, large-scale data analysis.

Some of these technologies may be familiar but still present barriers; others may be new or emerging. Technologies include:

- **Encryption**: This is a process by which content is rendered incomprehensible to everyone except specified receivers. Systematic data collection for researchers is impossible without access being granted by the sender or receiver.
- **New Formats**: Certain formats of content or data, particularly audio/audio-visual, are not (yet) as amenable to systematic search and storage as text. The nature of the content or data that researchers can gather and analyse from a platform has major implications for the kind of analysis that can be conducted. Textual data from traditional social media platforms like Facebook, Instagram, Twitter and VK can be relatively easily explored, especially where a systematic search function is available (e.g. through an API); however, predominantly audio-visual platforms like YouTube and Spotify present additional challenges as video and audio content cannot easily be searched or analysed in the same manner. Audio-visual content for AR/VR technologies is increasingly being developed and there is already evidence of it being used to spread harmful content or harass other users. This could substantially increase in the future if/when these types of technologies become more widely adopted.[10] The live and ephemeral nature of AR/VR activity also presents challenges for more systematic data collection approaches.
- **AI-generated content**: As demonstrated by "deep fakes", content generated by artificial intelligence is becoming increasingly believable. The speed at which new content can be developed makes systematic data collection harder.
- **Decentralisation**: This allows platforms to operate without central governance and can limit the ability of administrators to remove content or ban users (especially those users that have been identified as engaging in patterns of harmful behaviour). As well as decentralised platforms, there are projects aiming to allow decentralised communication between platforms.[vii] Decentralisation may therefore result in further fragmentation and reduce opportunities for more systemic data access for researchers.
- **Blockchain**: This is a technology via which events (e.g. who posted what content and when) are recorded on an unalterable ledger. This allows the current, true state of a system to be determined by consulting the current state of the ledger without the need for human intermediaries. Blockchain can therefore be used to accomplish decentralisation (e.g. platforms such as Riot). It is also often used to support payment in cryptocurrencies and, increasingly, platforms are using this to allow users to directly monetise content rather than relying on advertising (e.g. Odysee and LBRY). These financial incentives risk turning the pursuit of online harms into a business model, one which could prove particularly resistant to regulation or mitigation through its reliance on blockchain technology. From a research perspective, systematically collecting data from blockchain-based platforms remains relatively unexplored territory. As partially blockchain-based platforms like Odysee do not have public APIs, it is also unclear what data might become available and whether any further barriers might emerge during the process of data collection.

vii See the ecosystem review prepared before the launch of Twitter's decentralised protocol Bluesky.

Many of these technologies are barriers to specific forms of data access, and this may vary across platforms. For instance, consider an encrypted Telegram or Signal chat versus a private Facebook group. While both present similar ethical issues (as discussed shortly), for a researcher employing an ethnographic approach, Telegram or Signal's encryption is unlikely to present additional problems when compared with the unencrypted, private Facebook group; both require permission to be granted by the other users involved to gain access. But for the companies themselves, law enforcement or intelligence agencies, encryption presents additional technological barriers; the private and encrypted Telegram or Signal chat is inaccessible without permission from the users involved, whereas Meta could force access to data from the private but unencrypted Facebook group against the administrators' wishes.

Additional technological challenges to mitigating harmful content and behaviours

The range of these challenges is as great as the range of new technologies. Some examples might include:

- **New formats**: It is possible that new forms of content, perhaps AR/VR-based, will prove much more engaging and effective at radicalising audiences and/or helping harmful content achieve greater spread or impact. Market pressures may mean platforms would be unwilling to slow rollout of new technologies even in the face of such problems.

- **AI-generated content**: This could mean that content proliferates faster than it can be addressed. Automation (e.g. via "bots") is already used to rapidly duplicate and disseminate harmful content. More sophisticated AI could go beyond duplication, allowing material to mutate while retaining its original meaning.

- **Blockchain**: Particularly strict use of blockchain might make deletion of content by a centralised authority impossible or nearly impossible (e.g. a situation where an offending user would have to consent to the deletion of their content) though there are questions around how this would work in conjunction with legal requirements.[11]

Barrier Type 2: Ethical and Legal

Accessing data from online spaces, and particularly the collection and processing of that data, can raise ethical issues, such as invasions of privacy or the use of data or content without users' consent. This may also lead to contraventions of ethical research practices, platform terms and conditions, or even the law. This challenge can be particularly extreme for academic researchers who must often pass strict ethical approval procedures, as well as comply with relevant legal requirements. Law enforcement agencies (and intelligence services in many countries) are also subject to additional legal restrictions on their access to and use of personal data. This is desirable for a multitude of reasons, most notably the human right to privacy and ensuring due process. While the right to privacy is not absolute, exceptions need to be justified under the rule of law. Consequently, privacy restrictions can limit the ability to find harmful content. Some researchers have argued that the growth of privacy legislation across the world (most notably the General Data Protection Regulation (GDPR)[viii] in the EU and GDPR-influenced laws in other countries) may give platforms additional incentive not to share data.[12]

Messaging apps like WhatsApp are a pressing, current example. A huge amount of content is exchanged on WhatsApp, including forms of disinformation, incitements to violence and other harmful materials. If a researcher is a member of a WhatsApp group, collecting data is incredibly easy; WhatsApp has a simple functionality to export an entire chat history as a text file. But how did the researcher join said group? Did they gain explicit permission from all the members to use the group's content for research (potentially leading participants to self-censor)? Or are the group members unaware of the researcher in their chat, and therefore might they be non-consenting research participants? Did the researcher potentially gain access to the group via deception?

Similar ethical problems may arise when researching Discord. Discord's API client allows researchers to connect to a server and collect channel messages live as well as collect historic messages. There are two ways

viii GDPR is an EU data protection and privacy law. As it regulates the gather, storage and transfer of personal data, GDPR has important implications for online research.

in which researchers can connect to a server, each requiring different levels of identification or deception. In the first way, a bot account needs to be manually added to the server by a server admin (e.g. the server creator or someone else with those privileges); they may refuse such access. Additionally, the bot will be clearly identified as such in the user list which might raise suspicion, especially among communities discussing sensitive topics. The second way is to run a bot behind a regular user account (a so-called "self-bot"). In this case, researchers join the server as a normal user (e.g. with an invite link), and the bot subsequently impersonates this user; however, this deceptive behaviour contravenes Discord's Terms of Service (presenting an additional ethical challenge).

These problems may be even starker for messaging apps which, as a key part of their market offer, explicitly promise greater privacy and security than more mainstream options like WhatsApp. Platforms that promise a greater focus on the privacy of their users have also attracted harmful communities. For example, MeWe was founded in 2012 by privacy advocate Mark Weinstein and has since gained success among conspiracy theorists and far-right extremists.[13] Kik, an anonymous instant messaging service, has reportedly been used to facilitate child sexual exploitation.[14] As outlined in the above section on technological barriers, these platforms often use encryption. Additionally, such groups may be unlikely to welcome a potentially hostile researcher.

Additional ethical and legal challenges to mitigating harmful content and behaviours

As many platforms were created in response to increasing regulations and moderation practices in traditional social media, these alt-tech platforms are often presented as a bastion of "free speech" and therefore attract communities and ideologies that have been banned in other spaces for breaching community standards and/or hate, disinformation and harassment policies. This means platform moderation (and by extension terms and conditions, and general platform activity) may be explicitly opposed to actions such as content takedowns and banning accounts, or even downgrading harmful content in algorithmic recommendations, newsfeeds or search results.

For example, Minds explicitly markets itself as anti-censorship and pro-free speech. Although this attracts high levels of, for example, anti-vaccination content, moderators are highly unlikely to support taking it down (although the site does still take down illegal content). Nonetheless, the company still presents itself as opposed to misinformation, arguing that it should be fought via counterspeech.[ix]

Barrier Type 3: Fragmentation

Much online content, including harmful content, is theoretically accessible online without barriers caused by technological structures or ethical and legal issues; however, researchers still need to know where to look. Often relevant content is among vast amounts of material that cannot be searched quickly and systematically, for example, via a platform-wide search function or API. We refer to platforms where theoretically accessible content cannot be searched quickly or systematically as "fragmented". As the content is publicly visible, without technological or ethical and legal barriers, fragmented platforms may be seen as a subcategory of open platforms.[x] Not all open platforms will be fragmented, however, as some do offer the ability for researchers to systematically search content. Fragmented platforms are also distinct from closed platforms. While closed platforms also cannot be searched systematically, they cannot be accessed without additional information or permissions either (e.g. passwords or other types of personal identification).

This has been the situation for much of the history of research, as any historian who has ever needed to read through a poorly labelled physical archive can attest. Modern search tools (most notably Google but also platform-specific technologies like CrowdTangle[xi] or the Twitter API) have only recently increased the ease with which researchers can quickly and systematically locate content. This ease, however, can be (and often has been)

ix ISD defines counterspeech or counter-narratives as messages that offer a positive alternative to extremist propaganda and/or aim to deconstruct or delegitimise extremist narratives.

x While closed platforms cannot be searched systematically either, they also cannot be accessed without additional information (e.g. passwords or other types of personal identification). See Glossary.

xi CrowdTangle is a tool for searching public content on Facebook and Instagram. It is owned by Meta and over time, the company has limited the available data. Nonetheless, CrowdTangle still allows a quick keyword query to return an enormous range of material.

overstated. A huge amount of the web, potentially over 90%, does not appear in Google Search (this is the so-called "Deep Web").[xii] Furthermore, important forms of social media and online communication (private and/or encrypted messages, emails and closed groups) have always been off-limits to external researchers. Nonetheless, rapid and systematic searching has become vastly more possible as a technique for the discovery of harmful content and behaviour. But two converging trends may be reducing the power of these methods.

The first trend is that many online platforms, both new and established, are reducing the data that can be accessed through APIs or other tools. This is most notable on Facebook. Up to 2014, a researcher could use the "Graph Search" functionality of the Facebook API to access not just a user's data but also data about their friends. Even after substantial new limitations were announced in 2014,[15] researchers could still easily download every post on a public Facebook page and every comment made on that post, as well as associated profile information for each post and comment. In 2018, the API was greatly limited, and data access has since largely relied on going through Facebook partners (particularly CrowdTangle).[16] This means many key areas of platforms (e.g. private groups or pages) are beyond the scope of the API, forcing researchers to adopt older, more labour intensive and less systematic research methods, such as manually finding and reading material.

While increasing regulatory and public pressures have their benefits in terms of enhancing privacy and data rights, we may see that platform search tools and APIs become more restrictive by default. Alternatively, depending on the size of the platform and jurisdiction in which it operates, new regulations such as the EU's Digital Services Act may also result in more expansive access for researchers or the wider public. Many of the newer platforms in our case studies (see Section 4) do not have platform-wide search functions, even as part of their APIs. While it is still often possible to use relatively old technologies to access relevant data, this may involve more ad-hoc, "hacked together" methods, such as building a bot which mimics a human user and replicates text. These technologies need to be designed and maintained for specific purposes, including the production of data in a systematic format. This requires much more effort than using general search APIs. In some cases, using such technologies to access data may also break platforms' terms of service (ToS), thereby presenting additional ethical and legal challenges.

A second potential trend is the broader fragmentation of online hate spaces. The increasing willingness of many large platforms to claim they are acting against harmful content and behaviours may be driving these communities to seek (or build) a wide variety of alternative spaces. Particularly during the lead up to and right after the 6 January 2021 attack on the US Capitol, researchers saw pro-Trump actors moving from the increasingly inhospitable Facebook and Twitter to pre-existing "pro-free speech" spaces, such as Gab and Parler, which had allegedly also been used to coordinate parts of the insurrection.[17] In fact, Parler became the most downloaded app on 8 January 2021 after Facebook and Twitter suspended President Trump's accounts on their platforms.[18] After Parler was denied access to Amazon's cloud hosting service Amazon Web Services (or AWS) on 10 January 2021, users appeared to migrate to Gab. Over the next two months, 2.4 million accounts were created on Gab according to data leaks (Gab is believed to host some 4 million accounts though its active user base is estimated to be closer to 100,000).[19]

Technical features may contribute to this trend. Sites like nandbox allow users to easily create new messenger apps with little technical expertise. These types of service could facilitate the rapid fragmentation of potential spaces for hosting extremist content and communities. Helpfully for researchers, such platforms may provide obvious spaces to find and research harmful content if their popularity among extremists is widely known; however, we cannot assume that fragmentation will always lead to such obvious spaces to find harmful content. There is a range of large, fragmented platforms like Discord, Spotify and DLive on which harmful content could (and already does) go undetected amid a huge mass of other textual or audio-visual content.

Platforms can exhibit fragmentation in combination with other barriers. For instance, textual content may be

xii Technically, the Deep Web consists of online material which is not "indexed" by search engines and so will not appear in a search on Google, Bing, DuckDuckGo, etc. This includes a huge range of material that many people use daily, for example, any material which requires a password to access or is behind a paywall. The Deep Web is not to be confused with the "Dark Web", which can only be accessed through specific browsers and is often used for illegal activity.

openly accessible and distributed in comment threads under videos. However, without a systematic way to search audio-visual content (or capture and retain live-streamed content), the comments may present an incomplete picture of the relevant activity; this a combination of fragmentation and technological barriers. Alternatively, sites may mix private and public channels in such a way that it is unclear whether the entirety of relevant activity can be understood by extensive analysis of public channels alone. Access to private chats may also be required to fully understand the nature of the activity. This could require unacceptable levels of deception or participation to gain access, a combination of ethical and legal, and fragmentation barriers.

Additional fragmentation challenges to mitigating harmful content and behaviours

Even if harmful content and behaviours are discovered and addressed on one online platform, they can continue to proliferate across a variety of other platforms as users migrate across the online ecosystem. This is a long-standing issue in addressing harmful online activity, and some measures have been developed to address it. For instance, removal of illegal child abuse and terrorist content has used "hashing", giving images and videos a unique identifier so that replicas of a banned image can be more easily located (and again banned). Such hashing technology has been used by organisations like the Global Internet Forum to Counter Terrorism[20] and the Internet Watch Foundation.[21] Nevertheless, even with tools like this, complete removal of such content from the internet remains extremely challenging.

Moreover, if the precise form of the content varies or evolves (rather than being directly replicated), then tracking and removing similar or related content can be even harder. For instance, a specific copied image or video (e.g. the original livestreamed video footage of the Christchurch terrorist attack or the viral 'Plandemic' video[22]) would be easier to identify than edited versions of it or content that promoted a similar narrative (e.g. additional original content glorifying the Christchurch attack or promoting anti-vaccination disinformation). Here, the challenges to identifying relevant content posed by fragmentation may be further exacerbated if edited or similar content is spread at scale across a range of different platforms that cannot be searched quickly and systematically.

Section 3: Methodologies and Tools

Having laid out potential barriers, we now consider how methodologies commonly used by researchers into online spaces might respond to them. We begin by introducing three key types of method with reference to existing research and literature. We then cross-tabulate these with our three barriers to draw out the strengths and weaknesses of each method for addressing each barrier. There are very few instances where any of the barriers outlined in the previous section completely prevent research on a particular platform; however, they may severely limit the range of possible methods and tools that can be deployed. In addition to reviewing existing methodological approaches, we also conducted a scoping exercise to identify existing tools to find and collect content from alt-tech platforms. In the final part of this section, we present the findings from our scoping exercise, setting out the capabilities and limits of the analysis tools identified.

Method 1: Systematic Searching

This method involves using technology to extract large amounts of data and metadata directly from online platforms. Digital technologies have greatly increased the scale and ease of access to communications data. Various long-standing technologies, from copy-paste to web scraping, have allowed researchers to convert online data into easily analysed forms. Data might include, for example, the content of online text, connections between online accounts and metadata, such as times or geographical locations of posts.

The growing dominance of Web 2.0 platforms (designed to encourage user-generated content and participation), including social media platforms, has vastly expanded the range of this data. A researcher in the 2000s could track personal relationships by seeing, for example, how often different members of an online forum replied to one another. By the 2010s, researchers could see richer links of "friendship" between much larger audiences on platforms like MySpace or Facebook. Many social media platforms also made data easier to access by providing APIs; these have allowed researchers to directly access various forms of data from platforms without needing to build their own code from scratch.[xiii] The development of AI-based approaches has also allowed for ever more sophisticated analysis methods. For example, natural language processing (NLP) is increasingly used to detect trends, sentiments and entities mentioned across vast quantities of online text.

Much of the modern research into online platforms uses technological approaches to locate and collect data. The most popular tools include Google Search, the Twitter API or CrowdTangle (for Facebook and Instagram). External researchers have also developed other technologies. For example, CASM's Method52 allows for the collection and integration of data from multiple online platforms,[xiv] the mapping of relations between accounts and content, and the training of classifiers to distinguish researcher-defined themes within text. The Digital Methods Initiative (DMI) also provides a repository of tools developed by academics.[23]

The key advantages of systematic searching tools are:

- **Speed and scale**: Researchers can find, collect, and query billions of data points in seconds.
- **Systematicity**: While no tool provides an unbiased window into 100% of online data, the controllable and quantitative nature of these technologies allows for systematic collection and comparison (and potentially replication).
- **Precision**: A researcher skilled in querying techniques (e.g. Boolean operators) can focus a search onto precisely-defined content; AI-based technologies are increasing this capability still further. This is extremely valuable given the volume of online data researchers must frequently deal with.

The disadvantages are:

- **Data availability**: Research can become shaped by what data is available rather than by starting from a research problem and seeking the most appropriate data. Most notably, Twitter has received an outsized share of research attention relative to the size and

xiii APIs have also given platforms a greater degree of control over the data they supply, raising concerns around transparency and the stability of API-powered tools.

xiv Currently eight platforms, as well as external datasets, formats and sources (e.g. Media Cloud, Mastodon, RSS Feeds and Google Sheets).

diversity of its user base arguably due to the range of data it makes available to researchers in comparison with major platforms like Facebook, Instagram and especially TikTok.

- **Accuracy**: Research which relies on official APIs is dependent on the platforms providing continual access to accurate data. Platforms may not be incentivised to provide full and accurate data, and it is often hard to independently verify whether they are doing so. This issue also applies to datasets like Social Science One that have been compiled in collaboration with tech companies to provide access to external researchers. These have been beset by a series of challenges, including the accuracy of the data provided and the US-centric focus of the researchers granted access.[24] A reliance on companies to grant access for legitimate public interest research can also create disincentives for researchers to publicly criticise companies if their findings reveal failings in said companies' practices. Finally, it is sometimes possible for outsiders to create alternatives to APIs, but these may break platforms' terms and conditions and therefore expose researchers to potential legal risks.[xv]
- **Legal risks**: Third-party alternatives to APIs may break platforms' terms and conditions, thereby exposing researchers to potential legal risks.
- **Technical arms-races**: As online platforms increasingly diversify, incorporating ever more complex structures, metrics and types of media, it may become more difficult to develop tools which can access the full range of potentially relevant data and compare these across platforms. Researchers with the necessary financial resources and technological skills can outpace researchers who lack one or both, creating inequity within the research field and imbalances in the evidence-base.

Method 2: Ethnography

Ethnography is a well-established school of research methods which involves deep and sustained involvement with a community. Instead of relying on data-collection technologies, researchers may take a more human approach by joining, participating in and observing online spaces as forms of community.

Ethnography was a common approach in earlier research into online platforms, including many of the classic empirical works, for example, by Nancy Baym or Henry Jenkins.[25] This was accompanied by a growth in literature and research programmes on "digital anthropology" and "digital ethnography". While ethnography may now be less prominent than systematic search approaches, it is still a thriving research field.

The key advantages of ethnographic research methods are:

- **Contextual**: Ethnography can provide a rich, context-specific understanding of online activity.
- **Limited data**: It is suitable for studying niche subcultures that require immersion and do not produce the larger volumes of relevant data required for more quantitative approaches.
- **Alternate forms of content**: Ethnography research can involve the study of audio-visual content that cannot easily be analysed by technological tools commonly available to the researcher.
- **Vulnerability**: Ethnography is less vulnerable if platforms choose to clampdown on research tools (e.g. restricting data available through APIs).

The disadvantages are:

- **Hard to scale**: In-depth engagement with a community does not lend itself to the study of multiple platforms, and a human cannot parse as much data as technological tools.
- **Less systematic**: While ethnography may provide an in-depth understanding of specific communities, it does not provide a systematic view of wider online activity.
- **Ethical concerns**: Ethnographic research in closed spaces may require a degree of deception or impersonation, especially when researching secretive communities like violent extremist groups. Additionally, researchers may be directly exposed to harmful material or potential security risks.

xv See Annex: Legal Risks.

Method 3: Crowdsourcing and Surveying

Two less commonly used but potentially valuable methods for researching harmful content and behaviours are crowdsourcing and surveying. Crowdsourcing methods involve users of online platforms voluntarily reporting particular forms of content to researchers. Such reporting mechanisms can take multiple shapes like plug-ins[26] or reporting forms for users, offered either by third parties or online services themselves. A recent example is the use of "tiplines" for reporting dis/misinformation in WhatsApp chats during the 2019 Indian elections.[27]

At present, crowdsourcing methods are relatively novel, but their uptake on platforms like WhatsApp may encourage further attention. Harmful content voluntarily reported by users can also be used to create databases that assist in the research or prevention of malicious online activity. The GIFCT is a cross-platform initiative that maintains a hashing database containing fingerprints of known propaganda material by terrorist entities as designated by the United Nations.[28] Databases of violent content can also be used to preserve evidence for potential war crimes even if said content is removed from platforms for violating their policies. Such archives for collection and investigation exist for the wars in Syria[29] and Yemen.[30]

A related method for the voluntary reporting of harmful content is the surveying of internet users on their experiences. This approach has been deployed by national communications regulators like the Office of Communications (or Ofcom) in the UK.[31] In remote usability studies, users grant researchers access to their devices to monitor their digital behaviour. Such tests can be moderated, meaning all participants engage with the monitored service at the same time and can communicate with those conducting the research, or unmoderated, meaning users record their sessions at any time and send the recordings in later.[32]

Academics and research institutes have conducted similar surveys to research the effects of internet usage on users' attitudes and behaviours. De Zúñiga and Goyanes, for example, have used data from a two-wave panel survey in the US to argue that those who consume more news on WhatsApp (perhaps counter-intuitively) tend to know less about politics and are more likely to engage in unlawful political protest activities.[33] Researchers at the Centre for Monitoring, Analysis and Strategy (or CeMAS), an independent research organisation in Germany, have conducted survey research that establishes a correlation between the frequency of using the encrypted messaging platform Telegram (which is hugely popular among adherents of conspiracy theories) as a source of information and the readiness to protest against COVID-19 restrictions.[34] While this type of survey research is primarily focused on measuring the impact of internet usage, it can also be used in order to find out more about the reach of harmful content and narratives. In 2020, ISD supported a survey conducted by Tufts University about the prevalence of QAnon-related beliefs among the American population.[xvi]

The key advantages of crowdsourcing and surveying research methods are:

- **Combines advantages of systematic searching and ethnography**: Crowdsourcing and surveying methods find data via human participation rather than via platform-specific querying (and so are less vulnerable to, for example, API restrictions), but they also find data across a greater range of material than ethnographic methods.
- **Personalisation**: These research methods give insights into the personalised experiences of social media users. As algorithmic systems create different results based on a user's past behaviour, this approach allows researchers to gain insight into a wider range of user experiences.
- **Impact**: By allowing researchers to go beyond descriptively tracking online dynamics, they may be able to measure the impact of harmful content and behaviours online on wider political attitudes and behaviours. In particular, surveys are able to provide insights from audiences rather than just content-producers.

xvi QAnon is a wide-ranging conspiracy theory that claims an elite group of child-trafficking paedophiles have been ruling the world for decades. See 'Survey on QAnon and Conspiracy Beliefs', Tufts University and Institute for Strategic Dialogue, September 2020, https://www.isdglobal.org/wp-content/uploads/2020/10/qanon-and-conspiracy-beliefs-full_toplines.pdf.

The disadvantages are:

- **Accuracy**: As data is sourced from a variety of actors who may vary in diligence, understanding or levels of activity, it is hard to guarantee the systematicity, reliability and accuracy of inputs.
- **Sharing**: These research methods rely on group participants to share information outside the group. This may present ethical issues, and recruiting participants may be harder in certain groups (e.g. members of far-right groups may be unwilling to work with researchers who have been critical of the far-right).
- **Platform size limits**: It will likely be difficult to systematically survey users of smaller, more niche platforms given their smaller user bases, difficulties in identifying those that use these platforms, and their potential reluctance to participate in research.
- **Legal risks**: Certain crowdsourcing methods may present legal risks. For example, the use of third-party technologies (e.g. internet browser extensions or plug-ins) could contravene platforms' terms of service.[35]
- **Technological concerns**: It may require greater technical expertise and expense to create and operate technical tools, or to employ professional surveying companies.

Methods vs Barriers

The cross-tabulation below provides an overview of the applicability of each method in relation to each barrier, as well as any further issues.

Research Method	Technological Barriers	Ethical and Legal Barriers	Fragmentation Barriers
Systematic Searching	Widespread and continual monitoring can be used to discover early examples of emerging platforms and technologies. The technologies themselves could present barriers to large-scale systematic data access (see discussion in the fragmentation column).	Privacy and legal concerns are increasingly restricting the use of large-scale data collection without violating platforms' ToS.[xvii] There are ways to permit large-scale data access while preserving user privacy, for instance, "differential privacy" which introduces noise into the data to mask real identities. Many researchers are concerned that current techniques do not produce accurate results, particularly for research into specific content (e.g. harmful content). These techniques, however, are relatively new, and there is room for further development.[36]	Systematic searching has traditionally been the method used for addressing fragmentation barriers. Whether this continues to be the case will depend on the precise form of future platforms and searching/monitoring technologies. Increased fragmentation across niche platforms and/or loss of systematic API endpoints will limit the utility of systematic search technology. New developments in AI-powered search may enable systematic searching to adapt to these changes. Nonetheless, ethical problems with whether platforms permit this sort of data access could continue.
Ethnography	Potentially a powerful method against technological barriers; being part of a community allows the researcher to adapt to new technologies alongside other participants. May also give researchers early warning and insights into new technologies as they develop.	Deep, long-term involvement in a community may help ameliorate potential ethical concerns (e.g. participants may be more comfortable if they feel researchers are also community members). Conversely, deep, long-term involvement can also exacerbate ethical issues if, for example, a final report contravenes community expectations, researchers report detailed and personal information, or the research was based on a relationship of trust. For research into harmful content or behaviours, this negative scenario may be more likely.	Ethnography is unsuited to addressing this barrier; it is hard to scale and is generally unsuited to directly searching through large quantities of material. This is a trade-off against the deep and contextual understanding that is inherent to the method.
Crowd-sourcing	As demonstrated by ethnographic research methods, human participants can adapt to new technologies. They can also lead researchers to early examples of emerging technologies and platforms. Where possible, participants should be trained to help understand their understanding of relevant platforms and technological developments.	If crowdsourcing relies on existing participants of online communities, there are potential ethical grey areas around obtaining the informed consent of other participants that are not involved in or informed of the research; however, as long as sensitive personal data is not shared, crowdsourcing may be ethically justifiable. Participants' potentially poor understanding of privacy issues could lead to the over-sharing of data, resulting in ethical (and even legal) issues. If using "planted" participants, similar problems arise as for ethnography.	Large-scale crowdsourcing allows for a variety of platforms to be overseen by a variety of human monitors, and therefore may be well-placed to address issues of fragmentation. Issues of systematicity, reliability and scaling are present in such crowdsourcing.

xvii It should be noted that platforms may have other, more self-serving incentives for reducing data access. Limiting data access for researchers and journalists reduces transparency and therefore the risk of exposing platforms' failures to protect their users and wider society from online harms, as well as the role their products and business models can play in exacerbating or amplifying these harms.

Tools

The following section presents the findings from our scoping exercise aimed at identifying analysis tools for alt-tech platforms. While analysis tools dedicated to these platforms are rare, a few tools have been created over the years that allow systematic access to content and/or identification of broad metrics (e.g. followers, views and changes over time), or that can be used to support manual research efforts.

Social media analysis tools provide data access as well as the ability to monitor and analyse online conversations, trends and behaviours. These tools are widely used for a variety of purposes by marketing professionals, political parties, security services, government agencies and researchers.

Some tools may be freely available, but there are significant variations in levels of data access and transparency around the methods and technologies used to gather, analyse and present insights. Most of the tools that are widely used gather publicly available data from major social media platforms, such as Twitter (Brandwatch), Reddit (Pushshift), or Facebook and Instagram (CrowdTangle). But, while some tools allow keyboard-based searches of the entire platform, others limit searches to specific channels, accounts or communities of interest. CrowdTangle, which is widely used by journalists and researchers, does not provide systematic access to comments, just posts of public pages and groups. Broad trends data (e.g. tracking follower numbers, likes or video views) is available for most major platforms via open-source tools like Social Blade. This includes influential platforms like YouTube or TikTok that are often perceived to be difficult to research due to restricted data access (TikTok) or predominantly audio-visual content (both).

Potentially due to alt-tech platforms' more limited commercial importance and greater technical diversity, there are significantly fewer tools to monitor, track and analyse their content, trends and behaviours. Based on our review of existing literature that discusses those platforms identified during the platform-scoping exercise, ISD and CASM investigated thirteen tools that have some data access and, in certain cases, analytical features for alt-tech platforms: 4cat, Archived.Moe, Dewey Defend, DISBOARD, Lyzem, Method52, OSINT Combine Alt-Tech Social Search, Social Blade, Telegago, TelegramDB, Tgram.io, TGStat and Unfurl.[37]

Most of these tools do not provide systematic data access to the alt-tech platforms in question. Only 4cat, TGStat, Dewey Defend and Method52 allow some systematic access to content rather than just broad follower and view metrics, or account profile information. Furthermore, out of these, only 4cat is free and publicly available; it is an open-source analysis tool developed by Open Intelligence Lab (or OILab) and the DMI at the University of Amsterdam. As the name 4cat indicates, it specialises in gathering data from thread-based platforms like 4chan and, more recently, 8kun (formerly 8chan). It also allows researchers to create datasets from other platforms, including BitChute (scraping results from the video search function), Parler, Telegram (based on the researchers' Telegram API credentials) and Reddit (via the external Pushshift database). Based on the structure of the data acquired from each platform, further analysis modules are available within 4cat itself that allow, among other features, the identification of interrelated posts replying to each other, the detection of offensive speech and the collection of the most popular images within a dataset. 4cat is also relatively unique in having long-term data access to the chan-sites in particular; depending on the thread, data on 4chan can go back to 2012.

Other tools allow some systematic data access only to subscribers. For example, the public version of TGStat mainly provides access to broad metrics that show how subscriber numbers and views for Telegram channels have changed over time, but the ability to search for posts on Telegram containing keywords is limited to paying subscribers. Dewey Defend is similarly only accessible to licensed users, enabling them to find content on a wide range of platforms, including 4chan, 8kun, BitChute, Gab, Gettr, Kiwi Farms, MeWe, Minds, Parler, Poal, and Rumble, as well as Telegram channels manually added by users.

Beyond the few tools with systematic access to content and the tools which provide broad trends data, such as follower numbers, likes or video views (e.g. TGStat, and Social Blade which also covers Twitch, Odysee and DLive alongside major platforms), there is a range of tools that allows a researcher to search for specific content on different alt-tech platforms. For example, TelegramDB and Tgram.io allow users to search Telegram for groups,

channels and bots, while Telegago and Lyzem can additionally search through posts. With some of these tools, it is difficult to tell how the data is gathered and how comprehensive it is as the search results may appear incomplete (e.g. on Tgram.io). There are other search tools dedicated to single platforms, for example, Archived.Moe, where users can search all 4chan boards for posts (4cat restricts itself to selected boards like /pol/ or /k/), or DISBOARD, where users can search for Discord servers. Lastly, OSINT Combine, a company specialising in open-source intelligence training and software, has developed its Alt-Tech Social Search tool that enables users to search for posts on Parler, Gab, Minds, BitChute, DLive, Rumble, and several boards on JustPaste.it, WrongThink and 8kun. In terms of other alt-tech-specific, open-source intelligence tools, Unfurl extracts information from URLs, including timestamps and other domain information, and has a specific functionality to parse out information from Discord links.

Section 4: Platform Selection for Phase II Research

In the previous sections, we distinguished between three types of barriers to researching online platforms: technological, ethical and legal, and fragmentation (content that is public and accessible but not systematically searchable). These barriers are not mutually exclusive, and different functionalities within platforms may, at times, pose different barriers for researchers. Similarly, we identified three main methodological approaches to identify harmful content online: systematic search, ethnographic research, and crowdsourcing and surveying.

For the English, French and German language case studies that will be conducted during Phase II of this project, we have selected a combination of research barriers in addition to suitable methodological approaches to address these barriers. Based on the platform-scoping exercise, we have identified platforms that are increasingly used by harmful actors in each geographic context (one platform per context). Below, we outline some of the advantages and likely challenges of researching these platforms.

Fragmentation Barriers: Discord

For the case study looking at the United Kingdom, we propose researching harmful content and behaviour on Discord, a platform that primarily presents fragmentation barriers. In common with many other sites like Reddit and Facebook, Discord features a range of topical communities (or "servers") that users can join to chat with others. Many of these are private, but many are also public (though they still require a username to join). The largest public servers can have 100,000s of members[38], and while many public servers are dedicated to gaming or anime, some are dedicated to social/political discussion (some of which explicitly draw links to communities like 4chan).[39]

On Reddit, Facebook and other platforms, discussions in public groups can be accessed through the API. This means that a researcher can find mentions of relevant keywords (e.g. "Stop the Steal") quickly from across a range of public groups. A similarly wide-ranging functionality is not available on Discord; the capacity to search and download messages via Discord's API only functions server by server.[40] Some users have automated this to work at scale;[41] however, it appears that researchers would need to know in advance within which channels they wished to search. Given the huge range of channels on Discord and the fact that channels which host dubious content are sometimes deleted and/or renamed, this could make systematic searching very challenging.[xviii] The issue is not that the information is hidden; it would be easy to find if the researcher already knew where to look.

Discord's API client allows researchers to connect to a server and collect channel messages live as well as collect historic messages. In order to connect to a server, there are two ways in which researchers can identify themselves, both presenting technological and ethical challenges.

- **Bot account**: Per Discord rules, any automation needs to be run on a Bot account to prevent spamming, phishing and other malign behaviour.[42] Bot accounts cannot freely join servers; they need to be manually added to a server by a server admin (e.g. the server creator or someone else with those privileges), who may refuse this access. Additionally, the bot will be clearly identified as such in the user list which might raise suspicion, especially among communities discussing sensitive topics.
- **User account**: It is possible to run a bot behind a regular user account (a so-called *self-bot*). In this case, researchers join servers as a normal user (e.g. with an invite link), and the bot subsequently impersonates this user. This deceptive behaviour contravenes Discord's Terms of Service (presenting an additional ethical challenge). Discord may ban these accounts if they are discovered. It is unclear whether Discord is actively monitoring connections to discover accounts engaged in such deception or if they rely on them being reported by other users due to suspicious behaviour.

In addition, one of Discord's core functionalities is the combination of text and voice communication (e.g. so that gamers can play collaboratively and chat while they do so). Without the context of the voice call, much of the text communication may be uninformative.

xviii For instance, the Discord server "Slippy" (referenced in Levin, Nancy, '10 Largest Discord Servers', Largest.org, 18 August 2019, https://largest.org/technology/discord-servers/) appears to have been replaced with the server "Dream World" (https://discord.com/invite/dreamworld), though this is unclear.

Ethical Barriers: Telegram

For the case study looking at Germany, we propose researching harmful content and behaviour on **Telegram**, a platform that primarily presents ethical barriers. Telegram is a messenger app with platform-like qualities that has become a key online space for extremists, conspiracy theorists and disinformation actors. Particularly in Germany, Telegram has become a key hub for COVID-19-related conspiracy theories, disinformation and extremist mobilisation.

Telegram allows multiple modes of communication, including one-to-one messaging, group chats, private channels and public channels. Ethical (and, to some extent, technological) barriers for researchers therefore vary depending on the type of communication mode used on Telegram.

Public channels can have unlimited subscribers. While channel administrators may enable comment sections, they may also use Telegram exclusively for one-to-many communication. Therefore, bigger public channels present no particular ethical conundrums as there can hardly be a reasonable expectation of privacy, though it should be noted that the size of publicly visible Telegram channels may drastically vary, leading to expectations of privacy within small channels. Similar considerations apply to Telegram groups, which are limited to 200,000 members. Public groups may contain users with reasonable expectations of privacy, especially when they are smaller in size.

Despite its reputation as a privacy-focused platform, Telegram's API in fact provides data access for all channels and groups in which a user is registered. This includes historic data all the way back to when a channel or group was set up. Data acquired from groups will also contain some information about individual group members.

Access to content and group "membership" depends on the mode of communication. Content in public channels and groups is visible without joining; however, access to the systematic, historic data from public channels and groups is only be available to group or channel members. Joining may be as simple as merely clicking "Join", but there may also be further questions asked to screen admissions (and which may require deception on the part of researchers). Telegram limits the number of public and/or private groups and channels that one user (via a phone number) can join to 500, presenting some practical challenges around researching the platform.

One-to-one messaging and private groups on the other hand conform more closely to the description of Telegram as a messaging app like WhatsApp or Signal. These modes may also be used by the most extreme (and potentially violent) groups as a means of communication and mobilisation. Accessing these chats would likely not be possible without some level of deception.

Different methodologies could be used or even combined to research sub-sections of Telegram. A **systematic search** of links posted within public channels and groups could be used to identify potentially relevant closed chats (note that while Telegram provides an ID for the channels/groups from which content has been forwarded, these IDs cannot be used to automatically and systematically identify the names of relevant channels). **Ethnographic** methodologies could in turn be used to trial the feasibility of entering these closed (and likely high-risk) spaces within Telegram.

Technological Barriers: Odysee

For the case study looking at France, we propose researching harmful content and behaviour on **Odysee**, a platform that primarily presents technological barriers. Some platforms that are important to harmful actors may feature technological barriers which restrict access to data or have technical features that make researching them harder. Decentralised and/or blockchain-based platforms like Odysee present technological barriers worth investigating, especially in the context of an increasing presence of extremists and conspiracy theorists on the platform (particularly in France).

Odysee is a video-hosting platform that partially runs on LBRY, a decentralised blockchain-based file sharing network. Odysee was one of the platforms most frequently linked to in our datasets of French and German extremists, and it seems to be an increasingly popular, relatively libertarian alternative to those video-hosting platforms that have more stringent guidelines. This decentralisation makes it challenging to combat harmful content on Odysee as the technological ability of administrators to fully remove content (or records of content) and ban users may be limited.

In addition to being decentralised, Odysee is also blockchain-based and allows creators to monetise their content. Odysee offers the ability to monetise views (subject to, among other metrics, average watch time, average view count, type of content and engagement), direct donations and feature site/app promotions, all of which earn creators LBRY Credits.[43] After going through a cryptocurrency exchange, these can be turned into non-digital currencies.

As this is relatively unexplored territory, it would be worth testing if data from decentralised and/or blockchain-based platforms can be accessed systematically (**systematic search**), what data becomes available and whether, and so which, additional barriers arise during the process. As Odysee does not have a public API, it remains unclear whether it is possible to access data from the platform directly. Researchers would need to work on the LBRY network on which Odysee is built, and this could provide access to Odysee's video library, even though it remains to be seen whether comments and other metadata would be included. Such work may reveal that accessing useful data from the platform is impossible. It is possible that useful data can principally be accessed from Odysee, but that this would need major resources or require research methods that could be disputed from an ethical perspective. Hence, this work is in part aimed at simply identifying problems for researchers and practitioners that could become more urgent if Odysee continues to grow, especially among extremists and conspiracy theorists.

Section 5: Potential Future Scenarios

The previous sections have argued that technological developments, ethical considerations and issues of fragmentation may be increasing barriers to research of the broad ecosystem of online platforms. To exemplify how these trends could converge, we present two possible futures, one pessimistic and one optimistic. It is worth noting that the two scenarios outlined here are the extreme endpoints of a range of potential outcomes; the actual, future online ecosystem and regulatory environment may very well lie somewhere in between. The results will also vary across platforms, which already present a wide range of different functionalities, affordances, capabilities and corporate philosophies.[xix] We also provide a set of initial recommendations for policy-makers, regulators, researchers and platforms based on the findings of this report. These recommendations will be revisited and updated throughout the next phases of the project.

Pessimistic Scenario

A range of platforms develop which, either due to their ideological stance, business model and/or technical design, incubate harmful content and behaviours. They facilitate not just the growth of new narratives but also new technological developments, for instance, exploring how AR/VR can be used to create highly engaging radicalisation content or facilitate more visceral forms of online abuse and harassment, particularly targeting women, minorities and youth.[44]

These spaces are inaccessible unless researchers pretend to be members of extreme communities. An increasing range of screening technology is used to check identities, or researchers are required to demonstrate certain harmful behaviours before access is granted to an online space. Many researchers and, more crucially, ethics bodies are unwilling to support the levels of deception or participation required to join. The ratio of harmful activity to available researchers rapidly increases.

Through organisation and/or multi-platform integration, harmful content from these specialised spaces is able to quickly burst onto more mainstream platforms, thereby reaching new audiences and further amplifying harmful impacts. Blockchain-based monetisation of content encourages further spreading of the most engaging, radicalising or harmful content. Due to the widespread use of both AI and blockchain technology, once "in the wild", content can easily mutate and cannot easily be centrally controlled or effectively moderated. While mirror-image counter-hate spaces develop and attempt to use similar tactics and technology to the specialised hate spaces, these spaces find that they are consistently playing catch-up and their messages reach more limited audiences.

Additionally, platforms neither effectively address these problems, nor do they cooperate with researchers, and law enforcement or regulatory authorities. Regulation aimed at improving online safety, increasing transparency and providing regulators and researchers with access to data is ignored or resisted by certain platforms, especially those based in jurisdictions with weaker regulation, oversight or enforcement.[45] Smaller but highly toxic platforms that host harmful content or facilitate harmful behaviours fall through the cracks of laws that were primarily designed to regulate the largest and most dominant tech platforms.

Optimistic Scenario

The proliferation of platforms supposedly devoted to "free speech" leads to a fragmented landscape of spaces for harmful content, behaviours and communities. The increasingly niche nature of these spaces (different platforms for different kinds of hate, extremism and disinformation) allows specialised researchers to easily locate and identify harmful content and behaviours. While some of these platforms do place barriers on joining, these are not too onerous (to ensure new members are able to easily join). Continual marketing of new spaces means that relevant platforms are easily found by systematic monitoring, and intracommunal conflicts between groups and can also be leveraged to encourage leaking from private spaces frequented by hate, extremism or disinformation actors.

The current situation whereby narratives develop in specialised hate, extremism and disinformation spaces before spreading onto mainstream platforms continues; however, researchers are able, for the reasons outlined above, to prepare counter-methods against many

xix "Affordances" describe the technological opportunities provided to users by platform design and functionalities.

online harms in advance of them reaching and then being amplified in more mainstream spaces. Effective online regulations that outline clear transparency requirements and mechanisms for providing data access for research purposes are introduced and actively enforced. Platforms are willing to cooperate with researchers and regulatory authorities. Furthermore, the evolution of data protection and online safety laws leads to clear guidance and requirements on how to balance provision of data with privacy concerns. Developments in differential privacy allow researchers to access rich datasets without compromising personal privacy. The use of crowdsourcing methodologies (e.g. tiplines) also increases, aided by social media and messaging platforms that develop increasingly frictionless and engaging techniques for encouraging such behaviour.

Researchers and authorities are able to track a range of narratives as they develop through advances in AI, particularly:

- Increasingly powerful NLP, especially for audio-visual and live content formats.
- Self-generating data collection technologies that are able to train themselves to access the different platform structures that they encounter (and update themselves as these structures change).

Blockchain develops in a fashion that foregrounds transparency and accountability by default; this allows the source of harmful narratives to be more easily established.

Recommendations

Policy-makers and regulators:

- **When determining which platforms should be within the scope of regulation, policy-makers should consider the risks platforms pose, as well as their size, functionalities and number of users**. Where justified by higher levels of risk, governments should introduce appropriate and proportionate legal obligations on high-risk, smaller platforms to ensure that they do not become opaque online spaces dominated by harmful activity beyond the reach of regulators and researchers.
- **Policy-makers should ensure upcoming and future regulation includes sufficient platform transparency and data access provisions for regulators and approved external researchers**. In order to address technological and fragmentation barriers, platforms should be encouraged to take reasonable steps to provide structured and systematic data access. Where platforms are not within the scope of regulation that requires them to provide data access to researchers, policy-makers should introduce legal exemptions and/or protections for privacy-respecting, public-interest research to help build a greater understanding of the risks and harms on these platforms.
- **Policy-makers should consider how the regulation of social media platforms and other online services could be future-proofed** to account for the potential risks posed by a range of emerging technologies. Regulation should be designed with sufficient flexibility to allow regulators to adapt to new forms of harmful or illegal online activity, ensuring that regulation of the online ecosystem and its enforcement mitigates rather than simply displaces risks.
- **Policy-makers should ensure regulation incentivises and fosters "safety-by-design" approaches and ethical design principles across the technology sector** so that online risks and potential harms are considered in the design of new services, platforms or functionalities. Many of the platforms highlighted in this report have not been designed to facilitate harm, but there are instances where design changes could help to mitigate these risks. It is likely to be easier to consider these risks throughout the process of designing and launching a new platform, service or functionality rather than retrofitting mitigations in an attempt to offset fundamentally unsafe design choices.
- **Governments and regulators should cooperate with their counterparts internationally** to, as far as possible, avoid a divergent patchwork of online regulation. An inconsistent regulatory environment internationally would not only undermine the open, free and interoperable nature of the global internet, but it could also undermine attempts to make the internet safer by allowing companies and platforms to locate themselves in jurisdictions with the weakest regulation or no regulation at all. Governments and regulators should also coordinate to ensure consistency in requirements for data access; this would avoid over-burdening companies and forcing them to establish multiple, divergent processes and systems.

For researchers and civil society:

- **Civil society should continue to advocate for digital regulations that would protect and foster human rights online**. These regulations should strike an equitable balance between different rights, from freedom of expression through to privacy and protection from discrimination or incitement.
- **Civil society, academic researchers and funders of digital research should collaborate and invest in further developing research methods, tools and expertise** in order to keep pace with the rapid and continued evolution of the online ecosystem. New methods and tools will be vital to effectively monitoring and mapping this evolution as the diversity and applications of new technologies continue to grow (so too the range and types of risks posed by new or emerging platforms).
- **Civil society and academic researchers should continue to revise and harmonise existing norms, principles and guidelines for legal, ethical and secure online research.** This is particularly necessary for online spaces that are neither entirely public nor entirely private, and for emerging technologies like AR/VR. Researchers should also pool their resources and share expertise, including ethical guidelines, to address these increasingly complex legal, ethical and security challenges.

- **Civil society and academic researchers should develop shared, open repositories for recording and flagging potential platforms and/or technical developments of concern.** Certain platforms receive outsized levels of attention in social media research; there need to be crowd-sourced repositories and early warning systems which encompass more platforms across the online ecosystem. This should be done in a privacy-respecting fashion, for example, by not storing content or profile-level personal data.
- **As digital regulation is increasingly introduced in key jurisdictions, the research community and civil society should play a proactive role in helping companies and platforms to meet their regulatory compliance obligations and develop best practices**, especially those companies and platforms with more limited financial or technical resources, or limited expertise on the broad range of online risks and harms.

For platforms:

- **Companies should adopt "safety-by-design" approaches and ethical design principles when developing new online platforms and new features or functionalities for existing platforms**. These approaches encourage developers to consider throughout the design process the potential risks and impacts of new types of platforms, functionalities and emerging technologies, ultimately helping to ensure that mitigations are built-in rather than retrofitted. When developing new platforms or functionalities, companies should consult as early and widely as possible with civil society and academic experts on a broad range of online risks and harms, as well as with those impacted by them, particularly from disproportionately affected marginalised or minority communities.
- **Companies should permit public interest research in their platform's terms of service and be proactive in building constructive relationships with civil society and the research communities** to help identify, understand and mitigate potential risks and harms on their platforms. Platforms should also collaborate with each other to share best practice and identify emerging potential concerns and solutions.
- **Online platforms should provide access to public data via structured APIs and search functions, and (where possible) expand the scope of available public data while also respecting users' rights to privacy and security**. All areas of a platform that are public (and/or have a reasonable user expectation of visibility) plus all forms of content (i.e. textual and audio-visual content) hosted in these online spaces should be computationally transparent and accessible for privacy-respecting, public-interest research, including both near real-time and historic data. To the extent possible, data access should remain consistent so that long-term studies are not negatively impacted by changes or limitations in access.

Conclusion

As outlined in the previous section, Phase II of this project will conduct applied research to trial different methodological approaches across three platforms in an attempt to overcome the various barriers we have identified in this report and expand the research field's understanding of which methodologies are applicable with existing data access. Piloting new approaches will also allow us to reflect on the three types of barriers to research identified here (technological, ethical and legal, and fragmentation) and further update or augment them if required. These case studies, alongside the report's scoping of platforms, methods and tools, will be used to inform an assessment of the path forward for building practical solutions to access and analyse the ever-increasing and diverse range of online platforms.

The lessons learned from this research will feed into Phase III of the project, which will seek to inform practical, technical, and regulatory solutions to data access and transparency for these types of online spaces without impinging on the rights of users. We will share and discuss our findings with relevant research experts and technology company representatives whose work touches on data provision and transparency. We also hope to spark a wider conversation with other researchers so that they might provide recommendations and lessons learned from their own experiences of addressing the barriers we have identified, as well as the limits and obstacles involved. During this phase of the project, we will also engage with governments and policy-makers to share our findings on the evolving online ecosystem; the challenges, threats and opportunities this ongoing evolution presents for data access and transparency; and the implications for online safety, and regulatory and non-regulatory approaches to digital policy.

Finally, throughout the upcoming phases of the project, we will also revisit the potential scenarios and recommendations outlined above to further assess the range of future possibilities for the online ecosystem and regulatory environment as well as how researchers, policy-makers and platforms should respond to these changes. We will factor into these scenarios the findings and lessons learned from our upcoming research; inputs from other researchers and experts within the policy and privacy sectors; and any further technological or regulatory developments as they emerge. Too often over the past decade, digital researchers and policy-makers have struggled to keep pace with the rapid and vast changes that we have observed online, as well as the impacts that these have had on our rights, societies and democracies. We hope this project can provide a forward-looking contribution and ensure we are better prepared for what is to come.

Endnotes

2 'Questions and Answers: Digital Services Act', European Commission, 20 May 2022, https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348.

2 'The Draft Online Safety Bill and the legal but harmful debate', UK Parliament, 24 January 2022, https://publications.parliament.uk/pa/cm5802/cmselect/cmcumeds/1039/report.html.

3 'Community Standards', Facebook, https://transparency.fb.com/en-gb/policies/community-standards/; 'How to Use WhatsApp Responsibly', WhatsApp, https://faq.whatsapp.com/general/security-and-privacy/how-to-use-whatsapp-responsibly/; 'Community Guidelines', Instagram, https://www.facebook.com/help/instagram/477434105621119; 'Community Guidelines', Google, https://about.google/community-guidelines/; 'Community Guidelines', YouTube, https://www.youtube.com/intl/ALL_uk/howyoutubeworks/policies/community-guidelines/; 'Rules', Twitter, https://help.twitter.com/en/rules-and-policies/twitter-rules; 'Community Guidelines', TikTok, https://www.tiktok.com/community-guidelines; 'Code of Conduct', Microsoft, https://answers.microsoft.com/en-us/page/codeofconduct. For an overview of how these have evolved over time on Facebook, Instagram, Twitter and YouTube, see Katzenbach, Christian et al, The Platform Governance Archive, Alexander von Humboldt Institute for Internet and Society, 2021, https://doi.org/10.17605/OSF.IO/XSBPT.

4 Scrivens, Ryan et al, 'Examining Online Indicators of Extremism in Violent Right-Wing Extremist Forums', Studies in Conflict & Terrorism, 2021, https://doi.org/10.1080/1057610X.2021.1913818.

5 Goldman, Eric, 'Content Moderation Remedies', 28 Michigan Technology Law Review 1, Santa Clara Univ. Legal Studies Research Paper, 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3810580#.

6 Kreißel, Philip et al, 'Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz', Institute for Strategic Dialogue and Ichbinhier, 2018, https://www.isdglobal.org/wp-content/uploads/2018/07/ISD_Ich_Bin_Hier_2.pdf.

7 Guerin, Cécile and Fourel, Zoé, 'COVID-19: aperçu de la défiance anti-vaccinale sur les réseaux sociaux', Institute for Strategic Dialogue, 2021, https://www.isdglobal.org/wp-content/uploads/2021/04/COVID-19-aperçu-de-la-défiance-anti-vaccinale-sur-les-réseaux-sociaux.pdf.

8 O'Connor, Ciarán, 'The Conspiracy Consortium: Examining Discussions of COVID-19 Among Right-Wing Extermist Telegram Channels', Institute for Strategic Dialogue, 2021, https://www.isdglobal.org/wp-content/uploads/2021/12/The-Conspiracy-Consortium.pdf.

9 Gerster, Lea et al, 'Stützpfeiler Telegram. Wie Rechtsextreme und Verschwörungsideolog:innen auf Telegram ihre Infrastruktur ausbauen', Institute for Strategic Dialogue, 2021, https://www.isdglobal.org/wp-content/uploads/2021/12/ISD-Germany_Telegram.pdf.

10 For examples of documented harassment and abuse, see Basu, Tanya, 'The Metaverse has a groping problem already', MIT Technology Review, 16 December 2021, https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/; Bokinni, Yinka, 'A barrage of assault, racism and rape jokes: my nightmare trip into the metaverse', The Guardian, 25 April 2022, https://www.theguardian.com/tv-and-radio/2022/apr/25/a-barrage-of-assault-racism-and-jokes-my-nightmare-trip-into-the-metaverse; Robertson, Derek, 'Crimefighting in the Metaverse', Politico, 13 April 2022, https://www.politico.com/newsletters/digital-future-daily/2022/04/13/who-will-protect-you-in-the-metaverse-00025070. For examples of initial company research and responses, see Blackwell, Lindsay et al, 'Harassment in Social Virtual Reality: Challenges for Platform Governance', Proceedings of the ACM on Human-Computer Interaction, 3(100), November 2019, https://dl.acm.org/doi/10.1145/3359202; Gleason, Mike, 'Microsoft, Meta tackle harassment in virtual worlds', TechTarget, 17 February 2022, https://www.techtarget.com/searchunifiedcommunications/news/252513581/Microsoft-Meta-tackle-harassment-in-virtual-worlds.

11 Jurdak, Raja, Dorri, Ali and Kanhere, Salil S., 'Protecting the 'right to be forgotten' in the age of blockchain', The Conversation, 30 October 2018, https://theconversation.com/protecting-the-right-to-be-forgotten-in-the-age-of-blockchain-104847.

12 Shapiro, Elizabeth Hansen et al, 'New Approaches to Platform Data Research', Netgain Partnership, February 2021, https://drive.google.com/file/d/1bPsMbaBXAROUYVesaN3dCtfaZpXZgl0x/view.

13 Dickson, EJ, 'Inside MeWe, Where Anti-Vaxxers and Conspiracy Theorists Thrive', Rolling Stone, May 2019, https://www.rollingstone.com/culture/culture-features/mewe-anti-vaxxers-conspiracy-theorists-822746/.

14 Crawford, Angus, 'Kik chat app 'involved in 1,100 child abuse cases'', BBC, 21 September 2018, https://www.bbc.co.uk/news/uk-45568276.

15 Goel, Vindu, 'Facebook Promises Deeper Review of User Research, but Is Short on the Particulars', New York Times, 2 October 2014, https://www.nytimes.com/2014/10/03/technology/facebook-promises-a-deeper-review-of-its-user-research.html.

16 Perez, Sarah, 'Facebook rolls out more API restrictions and shutdowns', TechCrunch, 2 July 2018, https://tcrn.ch/2IKza9A.

17 Munn, Luke, 'More than a mob: Parler as preparatory media for the U.S. Capitol storming', First Monday, 26(3), February 2021, https://doi.org/10.5210/fm.v26i3.11574; Gais, Hannah and Cruz, Freddy, 'Far-Right Insurrectionists Organized Capitol Siege on Parler', SPLC, 8 January 2021, https://www.splcenter.org/hatewatch/2021/01/08/far-right-insurrectionists-organized-capitol-siege-parler.

18 Shieber, Jonathan, 'Parler jumps to No.1 on App Store after Facebook and Twitter ban Trump', TechCrunch, 9 January 2021, https://techcrunch.com/2021/01/09/parler-jumps-to-no-1-on-app-store-after-facebook-and-twitter-bans/.

19 Lee, Micah, 'Inside Gab, the Online Safe Space for Far-Right Extremists', The Intercept, 15 March 2021, https://theintercept.com/2021/03/15/gab-hack-donald-trump-parler-extremists/.

20 'FAQs / Explainers', Global Internet Forum to Counter Terrorism, https://gifct.org/explainers/.

21 'Image Hash List', Internet Watch Foundation, https://www.iwf.org.uk/our-technology/our-services/image-hash-list.

22 Macklin, Graham, 'The Christchurch Attacks: Livestream Terror in the Viral Video Age', Combating Terrorism Center, 12(6), July 2019, https://ctc.usma.edu/christchurch-attacks-livestream-terror-viral-video-age/; Frenkel, Sheera, Decker, Ben and Alba, Davey, 'How the 'Plandemic' Movie and Its Falsehoods Spread Widely Online', The New York Times, 21 May 2020, https://www.nytimes.com/2020/05/20/technology/plandemic-movie-youtube-facebook-coronavirus.html.

23 'Data Critique and Platform Dependencies: How to Study Social Media Data? Digital Methods Winder School and Data Sprint 2022', Digital Methods Initiative, https://wiki.digitalmethods.net/Dmi/WinterSchool2022.

24 Timberg, Craig, 'Facebook made big mistake in data it provided to researchers, undermining academic work', The Washington Post, 10 September 2021, https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/.

25 See particularly Baym, Nancy K., Tune In, Log On: Soaps Fandom, and Online Community, SAGE Publications, Inc., 2000; Jenkins, Henry, Convergence Culture, NYU Press, 2006.

26 'How it works', Ad Observer, https://adobserver.org.

27 Kazemi, Ashkan et al, 'Tiplines to Combat Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian Election on WhatsApp', arXic:2106.04726, July 2021, https://doi.org/10.48550/arXiv.2106.04726.

28 'Homepage', Global Internet Forum to Counter Terrorism, https://gifct.org/.

29 'Homepage', Syrian Archive, https://syrianarchive.org.

30 'Homepage', Yemeni Archive, https://yemeniarchive.org.

31 'User Experience of Potential Online Harms within Video Sharing Platforms', OFCOM (UK Government), 1 February 2020, https://www.gov.uk/find-digital-market-research/user-experience-of-potential-online-harms-within-video-sharing-platforms-ofcom.

32 Schade, Amy, 'Remote Usability Tests: Moderated and Unmoderated', Nielsen Norman Group, 12 October 2013, https://www.nngroup.com/articles/remote-usability-tests/.

33 Gil de Zúñiga, Homero and Goyanes, Manuel, 'Fueling civil disobedience in democracy: WhatsApp news use, political knowledge, and illegal political protest', New Media & Society, October 2021, https://doi.org/10.1177%2F14614448211047850.

34 Lamberty, Pia, Holnburger, Josef and Goedeke Tort, Maheba, 'CeMAS-Studie: Das Protestpotential während der COVID-19-Pandemie', CeMAS, 17 February 2022, https://cemas.io/blog/protestpotential/.

35 For example, see Bond, Shannon, 'NYU Researchers Were Studying Disinformation on Facebook. The Company Cut Them Off', NPR, 4 August 2021, https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f; Clark, Mike, 'Research Cannot Be the Justification for Compromising People's Privacy', Meta, 3 August 2021, https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/; Edelson, Laura and McCoy, Damon, 'We Research Misinformation on Facebook. It Just Disabled Our Accounts', The New York Times, 10 August 2021, https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html.

36 Shapiro et al, op. cit.

37 4cat, Archived.Moe, Dewey Defend, DISBOARD, Lyzem, Method52, OSINT Combine Alt-Tech Social Search, Social Blade, Telegago, TelegramDB, Tgram.io, TGStat and Unfurl.

38 'Top 100 Biggest Discord Servers', Discord, https://discords.com/servers/top-100.

39 'Discord servers tagged with 4chan', DISBOARD, https://disboard.org/servers/tag/4chan.

40 Discord API, https://discord.com/developers/docs/resources/channel#get-channel-messages

41 'Discord-Scraper', GitHub, https://github.com/Dracovian/Discord-Scraper#readme.

42 'Discord Developer Portal — Documentation — OAuth2', Discord, https://discord.com/developers/docs/topics/oauth2#bot-vs-user-accounts.

43 Leidig, Eviane, 'Odysee: The New YouTube for the Far-Right', Global Network on Extremism & Technology, 17 February 2021, https://gnet-research.org/2021/02/17/odysee-the-new-youtube-for-the-far-right/.

44 Bokinni, op. cit.

45 Meaker, Morgan, 'Germany Has Picked a Fight With Telegram', WIRED, 3 February 2022, https://www.wired.co.uk/article/germany-telegram-covid.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org