ISD | Powering solutions to extremism and polarisation

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

Executive Summary

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

Introduction 4
- Harmful Content and Behaviours Online 4

Finding Harmful Content and Behaviours 6

Platform Scoping: Methodology 7

Key Barriers to Online Research 8
- Barrier Type 1: Technological 8
- Barrier Type 2: Ethical and Legal 8
- Barrier Type 3: Fragmentation 9

Methodologies to Address Barriers to Online Research 11
- Method 1: Systematic Searching 11
- Method 2: Ethnography 11
- Method 3: Crowdsourcing and Surveying 12
- Methods vs. Barriers 14

Potential Future Scenarios 15
- Pessimistic Scenario 15
- Optimistic Scenario 15

Recommendations 17

Endnotes 19

Introduction

Recent decades have seen an important technological revolution: the increasing ability to systematically collect, store and precisely search communications data. 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.

Increasingly, various 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, and exploiting platforms that offer greater privacy, security or anonymity.

The full report considers these challenges in detail alongside the methods and tools currently available for researchers to monitor and analyse these types of platforms. It outlines findings from the first phase of a project that was funded by Omidyar Network and launched by the Institute for Strategic Dialogue (ISD) and CASM Technology. The accompanying annexes provide the full results of the platform-scoping exercise and further explore possible ethical, legal and security risks associated with researching these online platforms.

Phase II of the project will examine three platforms – Discord, Telegram and Odysee – in more depth to expand the research field's understanding of which methodologies are applicable to these types of online spaces within the bounds of existing data access. These case studies, alongside the scoping of platforms, methods and tools outlined in the report, 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 full report considers the barriers and challenges each platform presents in further detail.

During Phase III, ISD will seek to inform practical, technical and regulatory solutions to data access and transparency for these types of online spaces. We will 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. We will share and discuss our findings with relevant research experts and technology company representatives, and we hope to spark a wider conversation with other researchers. 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.

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, with how these changes have been exploited to cause harm, and with their impacts 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.

Harmful Content and Behaviours Online

Harmful content and behaviours can span a wide spectrum of online activity from 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.

Depending on the geographic and legal context, different forms of harmful content and behaviours may or may not be illegal. 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. 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.[1]

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 "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 fundamental 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 like Iron March or Fascist Forge).[2]

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 (e.g. CrowdTangle[i]); marketing-focused social media listening tools (e.g. Brandwatch); and research-focused technologies (e.g. Method52[ii]).

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 reach and, therefore, new audiences. 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.

Nevertheless, 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 (e.g. 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; secure communication technologies can 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.

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 online spaces that offer less moderation 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 French, German and English datasets focused on extremism or harmful conspiracy theories.

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

ii 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.

Platform Scoping: Methodology

In order to identity new and emerging platforms, ISD compiled a list of platforms and apps referred to by different harmful communities in 2021. To conduct this analysis, ISD and CASM Technology 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 extremism in French, German and English.[3]

Using these datasets, we were able to identify any links to other platforms shared in these groups. This exercise allowed us to list, in a systematic way, the most common platforms the communities in our study were linking to, and record and categorise barriers for finding harmful content on these platforms. The collection resulted in 35 platforms in French-speaking countries, 31 in German-speaking countries and 21 in English-speaking countries.[iii]

In order to narrow down this initial list of platforms and identify the most relevant for our research, we coded each platform across various categories, including:

- General information about each platform, such as the number of users, the purpose of the platform, and when and by whom it was founded
- Technological features (e.g. does the platform have a search function or Application Programming Interface (API[iv]); is it encrypted; and does it make use of new technologies, such as AR/VR or blockchain)
- Platform functionalities (e.g. does the platform offer closed[v] groups and private messaging)
- Whether the platform provides clear content policies, particularly around hate speech and disinformation
- Any terms and conditions relevant to data access and usage by external parties

Finally, barriers to research were also noted and categorised into three types (expanded on further in the following section):

- 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 scope of this exercise was ultimately to identify barriers to research, we restricted our final selection of platforms to those that present at least one of these three barriers. This exercise 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 both ethical (for closed groups) and fragmentation barriers (for 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.

iii See Annex: Platform-Scoping Data – Link Counts in the full report for the full list of each language.

iv An Application Programming Interface (API) is a software intermediary that allows two applications to communicate with each other.

v 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.

Key Barriers to Online Research

In this section, we present three types of barriers. 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 briefly introduce some of these challenges too.

Barrier Type 1: Technological

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.

Examples of these technologies and the additional challenges they present 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 forms of content or data are not (yet) as amenable to systematic search and storage. For example, primarily audio-visual platforms such as YouTube and Spotify present additional challenges because video and audio content cannot easily be searched or analysed in the same manner. AR/VR technologies are also increasingly being developed, and these could be used to spread harmful content or harass other users.[4] It may be 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. The live and ephemeral nature of AR/VR activity also presents challenges for more systematic data collection.
- **AI-generated content**: As demonstrated by "deep fakes", content generated by artificial intelligence is becoming increasingly believable. This could lead to content proliferating faster than it can be addressed. Additionally, more sophisticated AI could go beyond duplication, allowing content to mutate while retaining its original meaning. The speed at which new content can be developed also 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). Decentralisation may also 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 in 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). From a research perspective, systematically collecting data from blockchain-based platforms without public APIs remains relatively unexplored territory. 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).[5]

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) in the EU and GDPR-influenced laws in other countries) may give platforms additional incentive not to share data.[6]

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 material. 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?

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 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 become popular among conspiracy theorists and far-right extremists.[7] Kik, an anonymous instant messaging service, has reportedly been used to facilitate child sexual exploitation.[8] As outlined in the above section on technological barriers, these platforms often use encryption. Additionally, such groups are unlikely to welcome a potentially hostile researcher.

As many of these platforms were created in response to increasing regulations and moderation practices in traditional social media, these "alternative platforms" (or alt-tech[vi]) are often presented as bastions of "free speech" and therefore can 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.

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, one still does 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, fragmented platforms may be seen as a subcategory of open platforms.[vii] Not all open platforms will be fragmented, however, as some of them 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).

vi 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.

vii 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 Footnote V for a full definition of open and closed platforms.

Modern search tools (most notably Google but also platform-specific technologies like CrowdTangle[viii] or the Twitter API) have only recently increased the ease with which researchers could quickly and systematically locate content. This ease, however, can be (and often has been) overstated. A huge amount of the web, potentially over 90%, does not appear in Google Search (this is the so-called "Deep Web").[ix] 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 means many key areas of platforms 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. Many of the newer platforms identified in our scoping 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 and labour-intensive methods that need to be designed and maintained for specific purposes, including to produce data in a systematic format. In some cases, using such technologies to access data may also break platforms' terms of service, 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. Technical features may also 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. There is also a range of large, fragmented platforms like Discord, Spotify or DLive on which harmful content could (and already does) go undetected amid a huge mass of other textual or audio-visual content.

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 example, "hashing" to aid the removal of illegal child abuse and terrorist content.[9]

Nevertheless, even with tools like this, complete removal of such content from the internet remains extremely challenging. For example, 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. 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.

viii 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.

ix 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.

Methodologies to Address Barriers to Online Research

Here, we summarise three types of research methodologies for finding harmful online activity; their potential advantages and disadvantages; and how they can be employed to tackle each of the types of barriers outlined above. The full report also outlines the range of existing tools available to researchers for investigating harmful activity on smaller platforms.

Method 1: Systematic Searching

Systematic searching 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 online communications data, for example, the content of text, connections between 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. Many social media platforms have also made data easier to access by providing APIs. These allow researchers to directly access various forms of data from platforms without needing to build their own code from scratch.[x] The most popular tools include Google Search, the Twitter API, or CrowdTangle (for Facebook and Instagram). 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.

Advantages

- **Speed and scale**: Researchers can find, collect and query billions of data points in seconds.
- **Systematicity**: The controllable and quantitative nature of these technologies allows for systematic collection and comparison (and potentially replication).
- **Precision**: A researcher skilled in querying techniques can focus a search onto precisely defined content; AI-based technologies are increasing this capability still further. This is extremely valuable given the volumes of online data researchers must frequently deal with.

Disadvantages

- **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.
- **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. 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.
- **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-centric 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.[10] This was accompanied

x 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.

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.

Advantages

- **Contextual**: Ethnography can provide a rich, context-specific understanding of online activity.
- **Limited data**: It is suitable for studying niche online 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 technologies available to researchers.
- **Vulnerability**: It is less vulnerable if platforms choose to restrict research tools (e.g. restricting data available through APIs).

Disadvantages

- **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.

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[11] or reporting forms for users, offered either by third parties or online services themselves. At present these methods are relatively novel, but their use on platforms like WhatsApp may encourage further attention.[12] Harmful content voluntarily reported by users can also be used to create databases that assist in the research or prevention of malicious online activity, for example, to preserve evidence for potential war crimes even if said content is removed from platforms.[13] Surveying has also been used to understand the experiences of internet users both on- and offline. In remote usability studies, users grant researchers access to their devices to monitor their digital behaviour.

Advantages

- **Range of data**: These methods collect a wider range of data via human participation rather than platform-specific querying (making them less vulnerable to, for example, API restrictions). They also do so across a greater range of material than ethnographic methods.
- **Personalisation**: These methods provide 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 experience.
- **Impact**: Researchers can 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.

Disadvantages

- **Accuracy**: As data is sourced from a variety of actors, who may vary in diligence, understanding or levels of activity, it is difficult to guarantee the systematicity, reliability and accuracy of inputs.
- **Sharing**: These research methods rely on group participants sharing 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).

- **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.[14]

- **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.[xi] 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.[15]	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.

xi 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

Potential Future Scenarios

We 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.[xii]

Pessimistic Scenario

A range of platforms develop which, 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.[16]

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" the 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 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.[17] 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 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 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

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

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

Here, based on our findings, we provide a set of initial recommendations for policy-makers, regulators, researchers and platforms. These will be revisited and updated throughout the upcoming phases of this project.

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 our 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 protections 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 range of 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/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 community** 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.

Endnotes

1 '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.

2 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.

3 French: 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. German: 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. English: O'Connor, Ciarán, 'The Conspiracy Consortium: Examining Discussions of COVID-19 Among Right-Wing Extremist Telegram Channels', Institute for Strategic Dialogue, 2021, https://www.isdglobal.org/wp-content/uploads/2021/12/The-Conspiracy-Consortium.pdf. As the datasets were drawn from recent but distinct projects, the date range and sizes were varied. The French data included 2 million posts between 31 July 2020 and 31 January 2021. The German data included 659,000 posts between 1 January 2021 and 12 September 2021. The English data included 2.5 million posts between 1 January 2021 and 30 November 2021.

4 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.

5 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.

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

7 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/.

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

9 See 'FAQs / Explainers', Global Internet Forum to Counter Terrorism, https://gifct.org/explainers/; 'Image Hash List', Internet Watch Foundation, https://www.iwf.org.uk/our-technology/our-services/image-hash-list.

10 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.

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

12 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.

13 See 'Homepage', Syrian Archive, https://syrianarchive.org; 'Homepage', Yemeni Archive, https://yemeniarchive.org.

14 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.

15 Shapiro et al, op. cit.

16 Bokinni, op. cit.

17 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