EXTRACTS FROM ISD'S SUBMITTED RESPONSE TO THE UK GOVERNMENT ONLINE HARMS WHITE PAPER

July 2019

On 8th April 2019, the UK Government's Department for Culture, Media and Sport (DCMS) released its much-anticipated Online Harms White Paper, setting out a range of proposed legislative and non-legislative measures to tackle online harms ranging from illegal (e.g. terrorist content) through to harmful but not necessarily illegal online behaviour.

In response to a call for consultations on the White Paper, ISD provided a submission that was developed in collaboration with Demos, Luminate, Digital Action, and grounded in discussions held with a variety of other relevant civil society organisations.

*A core aspect of this submission focused on the questions of **transparency**, **public and private online spaces**, **safety by design** and **educating and empowering internet users**, as well as feedback on the proposed **regulatory scope, proportionality and oversight**.*

Transparency

Transparency is a commonly-used, 'catch-all' solution for improving visibility, understanding and accountability for online harms. By increasing transparency of online spaces, the argument goes, we stand a better chance of detecting, mitigating and responding to online harms. However, the requirements and expectations associated with transparency are often poorly articulated.

It is of central importance that governments, civil society and the public are able to better understand the ways in which the internet is impacting society and democracy in order to better to encourage its positive effects and to curb negative externalities. These kinds of decisions require evidence: transparency is the tool through which this evidence can be gathered. Good models for transparency exist, and should be used by the government as best practice when thinking about the wider online ecosystem and the future of expectations for transparency online.

This response aims to clarify what should be expected from online spaces with regards to transparency and to lay out a framework for transparency requirements. These recommendations are to be read alongside existing commitments in the White Paper to protections for freedom of speech and expression and proportional application of the duty of care.

We aim to support conceptual recommendations with clear technical guidance. We believe that good models for transparency exist, and should be used by the government as best

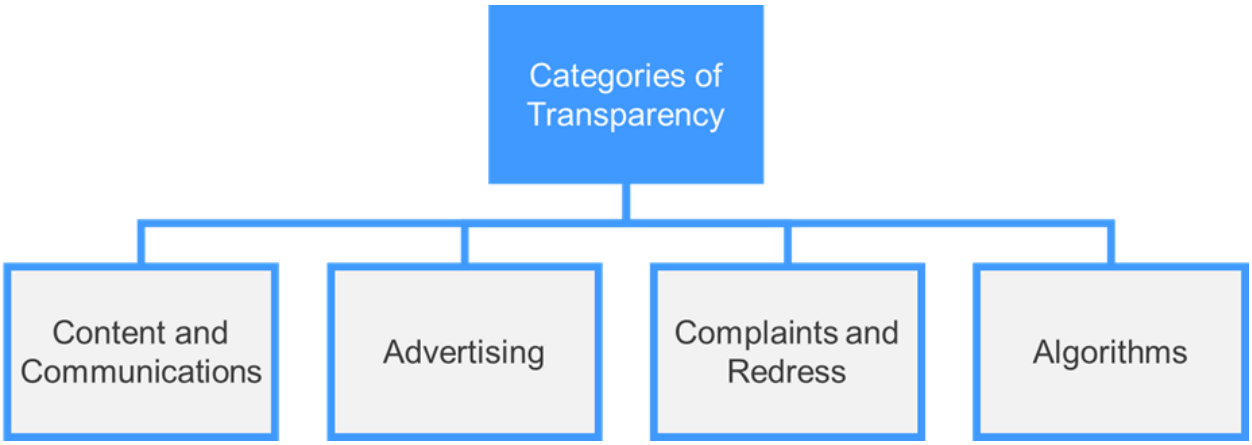
practice when thinking about the wider online ecosystem and the future of expectations for transparency online.

Why should platforms be transparent?

The migration of our lives onto the internet has created powerful new forces that shape citizens' lives. Online platforms have radically altered the ways in which we communicate, make decisions and come to hold views. Given their influence, it is of central importance that governments, civil society and the public are able to better understand the ways in which the internet is impacting on society, to encourage its positives and to curb its negatives. These kinds of decisions require evidence: transparency is the tool through which this evidence can be gathered.

Four Categories of Transparency

Going forward, we believe platforms should have transparency in four categories.



1. **Content and content moderation:** Platforms that have become public spaces must make that space as intelligible as possible. As web platforms and their users play an increasing role in shaping our culture, informing our political decision-making, and driving societal change, the activities taking place in these spaces should be observable.

2. **Advertising:** Advertising - particularly political advertising - has been shown to be a key vector through which publics can be manipulated. It is in the public interest for internet users to understand how and why they are being targeted online, and for regulators to be able to understand and respond to malpractice.[1]

[1] See the open letter signed by ISD and 10 expert organisations for an effective advertising API model, which stresses that all official communications by political parties should be accessible, at a minimum at the level of content and metadata.

3. **Complaints and redress:** A significant gap exists in the publics' understanding of platforms' ability to moderate and respond to abuses of their platforms. Visibility of complaints made to platforms is essential to accountability, to support the victims of online harms, to raise awareness of challenges facing users online and to provide evidence in redress.

4. **Algorithms:** There remains significant concern that platform architectures contribute to negative outcomes. Central to this concern is that the algorithms dictating a users' experience and journey have led to unintended consequences, and have been challenging to scrutinise or evaluate.

Principles of Transparency

Each of these areas requires a different set of frameworks. However, drawing them together are some broad principles of digital transparency.

1. **Transparency must be computational.** For an online space to be transparent, it must be possible to observe it computationally. For instance, Twitter's API allows for a holistic view of what takes place on that platform. Were that API not to exist, an otherwise nominally 'public' platform would not be transparent as a direct result of its scale - it overwhelms human capacity.

2. **Transparency must complement rights to data privacy, not erode them**. A good model for transparency will protect individuals' data privacy while enabling a macro understanding of the nature and scale of technology platforms' processes and any potential infringement of rights that stems from the use of the platform.

Data Protection and Risks

- Access to transparency data may be contentious. Although we believe that it is in the public interest to have oversight over all four areas, it is possible that there ought to be exceptions to the types of data and types of access available to the general public. A 'tiered' access structure (by which a regulator or institutions accredited by a regulator or other body have increased access to transparency data) may be advisable in light of data protection and privacy expectations. However, we believe the starting point ought to be public access.
- The government could use the model of data trusts, currently being piloted by the Office for AI and the Open Data Institute, to provide an independent body (alongside the regulator) who would determine access to company's data.

Transparency Category 1: Public Communications and Content

Overview

The public content and communications taking place on online platforms should be transparent. As noted above, we expect public communications that are subject to transparency to meet certain standards of transparency. Evaluating whether an online space is a public space is a challenge. In ascertaining how public or private a function of a platform is, we refer to the briefing paper one, *Public and Private Online*.

In short, those spaces judged to be public should be transparent.

Central to this requirement will be computational transparency. For services of a certain scale, this would likely require an Application Program Interface (API). We believe Twitter is an excellent model for this kind of transparency: those functions of the platform that are public, and have a reasonable user expectation of publicity, are computationally transparent and accessible through a well-supported API. By contrast, decisions to depreciate API access to certain public parts of Facebook, such as public Facebook pages, have moved nominally public spaces beyond observation.

We recommend that public online spaces should be expected to provide the mechanisms to observe them. In line with commitments to proportionality, platforms judged to play a major role in providing public spaces should provide a documented API (or similar).

Technical Recommendations

We set out the following requirements for an API to be deemed fit for purpose in providing computational transparency to an online space. We draw heavily on Mozilla's recent recommendations regarding a functional advertising API, copied out in full in area 3 below. Each section is illustrated with an example taken from Twitter, which at the time of writing we believe to be the gold standard.

A public content and communications API should have the following:

1. Full Content Access

- The APIs should include all content circulating in a public space, searchable by public identifiers.
    - Content includes text, media and links with associated metadata including public account information and timestamps.
    - Public identifiers includes the keywords of textual content, the URL of a link or piece of media, or the handle used by an account.

In the case of Twitter, researchers are currently able to use its Search and Stream APIs to collect data based on users or groups of users, keywords, or time periods.[2]

2. Live and historical data

- APIs should support analysis that looks to monitor public spaces in real time, and to support analysis that examines change over time by providing data over longer historical periods.

Twitter's standard Stream API provides a high degree of transparency in real time. However, Twitter's public-access search function is currently limited to seven days of historical data. Access to the full archive of Twitter data is available, but is paid for. We believe this period should be increased to thirty days, and believe there are arguments for making the full archive of a public space available for observation.

3. Register of use

- Institutions or users collecting data from an API should do so publicly, with a publicly available record of who has accessed data through an API and details of their search queries.

It is our understanding that social media platforms keep internal records of API use. These should be public.

4. Functionality to empower, not limit, research and analysis, including:

- All images, videos, and other content in a machine-readable format accessible via a programmatic interface.
- The ability to download a week's worth of data in less than six hours and a day's worth of data in less than one hour.
- Bulk downloading functionality of all relevant content. It should be feasible to download all historical data within one week.
- Search functionality by the text of the content itself, by the content author or by date range.

Twitter's APIs output in a standard machine-readable format (JSON) with generous rate limits at no cost. We believe this is further evidence of a high-quality API.

[2] Twitter's Search API is documented at https://developer.twitter.com/en/docs/tweets/search/overview and its Stream API is documented at https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

5. Public access.

- The API should be accessible to the general public, subject to the discussion of risks above.

Transparency Category 2: Advertising

Overview

Recent years have seen an explosion in data which can be used to target advertising, including location, IP, browsing data, and information collected from devices and wearables. The next few years are likely to see a move to increasingly automated marketing, where valuable individuals or groups are tracked, measured and targeted by machines, potentially using machine-generated content. There are major risks associated with the use of this kind of data for advertising.

This risk is most prominent in political advertising. Elections are becoming increasingly 'datafied', with advertising and marketing techniques being offered by a network of private contractors and data analysts, offering cutting-edge methods for audience segmentation and targeting to political parties all over the world.[3] Questions of user consent, knowledge and privacy have not been answered, and accountability and visibility of this kind of advertising is not adequate.

Recent efforts by platforms to increase transparency around advertising have been inadequate. Statements by the European Commission earlier this year have underlined the need for further effort by platforms to make advertising data available to scrutiny. We recommend online platforms are expected to provide functional, computational transparency for advertising. Mozilla's recent letter to social media platforms outlining the requirements of a functional advertising API in light of suspected exploitation and abuse of digital advertising provides a highly useful blueprint for advertising APIs.

Technical Recommendations

As noted above, the following expectations of an advertising API were compiled and circulated by Mozilla in April 2019, and the accompanying letter was undersigned by over seventy independent researchers in this space.[4] A functional, open API should have the following:

1. Comprehensive political advertising content.

- The APIs should include paid political ads and issue-based ads, without limiting access on the basis of pre-selected topics or keywords. "Political" ads might include, but are not limited to:

[3] https://demosuk.wpengine.com/wp-content/uploads/2018/07/The-Future-of-Political-Campaigning.pdf
[4] blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/

- direct electioneering content
- candidates or holders of political office
- matters of legislation or decisions of a court
- functions of government

2. The content of the advertisement and information about targeting criteria, including:

- The text, image, and/or video content and information about where the ad appeared (newsfeed, sidebar, etc.).
- The targeting criteria used by advertisers to design their ad campaign, as well as information about the audience that the ad actually reached.
- The number of impressions that an ad received within specific geographic and demographic criteria (e.g. within a political district, in a certain age range), broken down by paid vs. organic reach.
- The amount of engagements that an ad received.
- Information about how much an advertiser paid to place the ad.
- Information about additional platform functionality, including whether the ad was a/b tested and the different versions of the ad, if the ad used a lookalike audience, the features (age, gender, geographic, etc.) used to create that audience, if the ad was directed at platform defined user segments or interests, the segments or interests used, or if the ad was targeted based on a user list the advertiser already possessed.

3. Functionality to empower, not limit, research and analysis, including:

- Unique identifiers associated with each advertisement and advertiser to allow for trend analysis over time and across platforms.
- All images, videos, and other content in a machine-readable format accessible via a programmatic interface.
- The ability to download a week's worth of data in less than six hours and a day's worth of data in less than one hour.
- Bulk downloading functionality of all relevant content. It should be feasible to download all historical data within one week.
- Search functionality by the text of the content itself, by the content author or by date range.

4. Up-to-date and historical data access, including:

- Availability of advertisements within 24 hours of publication.
- Availability of advertisements going back 10 years

- APIs should be promptly fixed when they are broken, and to the extent possible remain consistent so that long-term studies are not negatively impacted by changes to APIs.[5]

5. Public access:

- The API itself and any data collected from the API should be accessible to and shareable with the general public.

Transparency Category 3: Complaints and Redress

Overview

A further category of transparency we believe would work to improve the health of platforms is over complaints and reports made to platforms and platforms' responses. Collective redress would encourage UK citizens to defend their rights online: for this to happen, it is vital the machinery is in place to draw attention to malpractice, abuse and failures in protection on platforms.

In matters of disinformation and large-scale exploitation or manipulation of online spaces, computational transparency over the content and communications as outlined in Category One above should act as a system to alert authorities to potential breaches. Transparency into user complaints adds a further layer of protection by ensuring that individual complaints can be understood at scale.

Moreover, there remains significant controversy as to the efficacy of platforms in responding to users' complaints, with frequent reports of inaction in the face of clear breaches to terms of service, and risks to successful redress as a result of content removal in response to these breaches. Research groups have begun testing the response of social media platforms to complaints by measuring response times and outcomes, though this work is laborious and poorly supported by current platform architectures.

We recommend platforms commit to archiving complaints in a machine-readable archive, allowing third-party oversight of issues impacting web platforms' users and the companies' record in responding to their complaints. There is precedent for this, with public bodies in the UK publishing complaint-level data (for instance, NHS CCGs).[6] We anticipate that access to this data would be limited to institutions approved by the regulator given its sensitivity.

[5] This may require setting out what sufficient notice of forthcoming changes are and reasonable transition periods in the event of API change.

[6] https://www.ashfordccg.nhs.uk/contact-us/comments-and-complaints/complaints-archive/

Technical Recommendations

At this stage, there is very little information as to how complaints might be used by regulators, ombudsmen or support groups to improve users' experiences online. We propose below a series of considerations we believe might form the partial basis of a useful dataset, but stress the need for further research and development here in partnership with these groups.

1. Comprehensive complaints data

- Anonymised complaints data allowing reviewers to review complaints at scale without compromising the privacy of complainants or those being complained about.
- Metadata, including timestamps.
- A record of the time taken to respond and resolve a complaint.
- A record of the final decision made, including a record of whether content was removed.

2. Protections for users

- Complaints data should be anonymised to prevent reidentification of complainants.

3. Functionality to empower, not limit, research and analysis, including:

- Delivery in a standardised machine-readable format.
- The ability to download a week's worth of data in less than six hours and a day's worth of data in less than one hour.
- Bulk downloading functionality of all relevant content. It should be feasible to download all historical data within one week.
- Search functionality by the text of the content itself, by a consistent but anonymised author ID or by date range.

4. Up-to-date and historical data access, including:

- Availability of complaints data within 24 hours of publication.
- Availability of complaints data going back 10 years

Transparency Category 4: Algorithms

Overview

There is a need to establish new systems for oversight or auditing of algorithmic decision-making. The crucial role which algorithms play within how social media platforms, and the types of problems they may contribute to such as the promotion of hate speech or polarizing an already divisive political culture requires as such.

Existing efforts in this area already underway, such as those between academia and the private sector to allow external researchers to analyse information amassed by companies to address societal issues are of course important.[7] These efforts have been challenging to set up, have yet to prove themselves, and are limited in scope. But it is right that the online harm regulator will adopt a role to "encourage and oversee the fulfilment of companies' commitments to improve the ability of independent researchers to access their data".

The Online Harms white paper also sets out that "[w]here necessary, to establish that companies are adequately fulfilling the duty of care, the regulator will have the power to request explanations about the way algorithms operate."

In order to give effect to the stated aims of the white paper and to ensure true algorithmic accountability, we feel the regulator should have powers to undertake algorithmic audits themselves.

To do this, the regulator should be able to examine the purpose, constitution, and policies of the systems and to interview people who build and interact with different parts of that system, and observe how people use the system. The regulator should be able to identify and assess what data was used to train the algorithm, how it was collected, and whether it is enriched with other data sources, and whether that data changed over time. It should be able to examine the model itself including considering the processing flow and the type of supervisory or monitoring mechanism used. The regulator should be able to undertake a code review or "white-box testing" to analyse the source code, or the statistical models in use, including how different inputs are weighted.

The regulator should also be empowered to run controlled experiments over time to determine if the algorithms subject to their review are producing unintended consequences that harm the public interest. Such experiments would be novel in this area, but such independent testing and experimentation are of course commonplace in other areas such as pharmaceuticals or food safety.

It is only by undertaking this sort of audit will the regulator, acting in the public interest, be able to assess whether companies truly are acting responsibly, protecting the safety of their users and tackling harms on their platforms.

In order to achieve this, the staffing of the online harms regulator would need to include a number of technologists who are able to advise to develop policies and procedures on how such an audit should take place as well as undertaking the audit. The regulator could also bring in external experts as consultants or fellows to assist with particularly complex or novel issues, as other regulators have in the past.[8] This work should draw on efforts already

[7] https://socialscience.one/

[8] See Interception of Communications Commissioner's report into the use of Chapter 2 of Part 1 of the Regulation of Investigatory Powers Act to identify journalistic sources which was assisted by Professor Anne Flanagan, Professor of Communications Law at Queen Mary University of London.

underway by the Information Commissioner's Office to create an AI auditing framework for data protection.[9]

This work should be done openly and transparently, with continuing consultation with civil society, industry and academia to jointly pool expertise to establish best practice in this new area. It is envisaged that processes could be co-constructed with external stakeholders and that this approach would be continuously developing, and responsive to the changing technological landscape.

There are a couple of different models that could be drawn from in terms of powers of audit. The ICO can undertake consensual audits to carry out an assessment of data controllers or processors are complying with good practice in the processing of personal data[10]. Should the company not agree to a consensual audit, the ICO can (should they decide that there are reasonable grounds for suspecting a data controller or processor is failing to comply with the Data Protection Act) seek a warrant to enter, search, inspect, examine and operate any equipment in order to determine whether a company is complying with the act.[11]

A similar power to be provided to the online harms regulator, to empower them with the consent of the company, to carry out an algorithmic audit, or if the company doesn't provide consent, and there are reasonable grounds to suspect they are failing to comply with the duty of care, to seek a warrant to determine whether they are.

Alternatively, a model could be drawn on that is used by the Investigatory Powers Commissioners Office (IPCO) who are responsible for keeping under review the use of investigatory powers by a number of public authorities including the security and intelligence agencies and law enforcement bodies. IPCO has powers[12] to conduct investigations, inspections and audits as the Commissioner considers appropriate for the purpose of the Commissioner's functions including access to apparatus, systems or other facilities or services.[13] In practice, this means IPCO are able to inspect on site, the entire system used by the body they are auditing, including the underlying data, any technologies processing the data, and the output provided.

It is, of course, important that commercial confidentiality is respected, as well as the data protection of the data being used by the algorithm. To avoid unnecessary risk, and to protect commercially confidential information and to ensure the protection of personal data, it is envisaged that the presumption would be that such audits would be undertaken entirely on the companies' systems within their premises or alternative secure location, and not that data, or commercially confidential information, would be removed offsite for analysis.

[9] https://ai-auditingframework.blogspot.com/
[10] s129, Part 5, Data Protection Act 2018
[11] Schedule 15, Data Protection Act 2018
[12] s235(1), Chapter 1, Part 8, Investigatory Powers Act 2016
[13] s235(4), Chapter 1, Part 8, Investigatory Powers Act 2016

Defining Public and Private Spaces Online

The legal status of online spaces as private or public has been reopened by the White Paper. The paper states that "the framework will ensure a differentiated approach for private communication, meaning any requirements to scan or monitor content for tightly defined categories will not apply to private channels".

We believe that the continuing debate over the legality and ability of authorities to oversee aspects of online life to be reflection of confusion as to whether certain spaces online constitute public or private spaces. Private life and communications have been interpreted broadly by the European Court of Human Rights. The Court has highlighted that "private life is a broad term not susceptible to exhaustive definition". This lack of exhaustive definition has been reflected in conversations we have had with law enforcement, parliamentarians and government officials who have also requested guidance as to when online spaces should be treated as public or private.

A handful of cases have raised the question, attempting to determine the significance of expectations of privacy amid current legal precedents. Emerging regulatory frameworks are now also challenged with the question of whether online platforms can be defined as private or public, or whether we require a broader, non-binary spectrum of public, semi-private and private definitions for the emerging world of online communications systems.[14]

Meanwhile, decisions by platforms to reduce the ability of civil society organisations and the public to observe, analyse and report on what is taking place online are troubling. Web platforms, and social media platforms in particular, are essential sources of insight on the ways in which the web is changing politics, society and culture, including the impact potential online harms.

We believe that civil society research organisations, academia, open source intelligence groups and investigative journalists have a vital role to play here, both in using online sources to report on changes in the online world and in evaluating the roles and decisions made by online platforms. The work done by Bellingcat that found responsibility for the downing of MH17 lay with Russian separatists, for instance, would have been impossible without access to online data sources.

There are dozens of research organisations working to better understand how the internet is shaping our lives. It is vital the White Paper protects and encourages their work.

[14] Recent cases include Facebook Inc. Consumer Privacy User Profile Litigation, California (2019), http://www.documentcloud.org/documents/6153329-05-29-2019-Facebook-Inc-Consumer-Privacy.html with reporting: https://www.law360.com/articles/1164091/facebook-says-social-media-users-can-t-expect-privacy and https://theintercept.com/2019/06/14/facebook-privacy-policy-court/,

Public and Private

Our focus is twofold: to propose possible tests for determining whether an online space should or should not be accessible to civil society organisations and research groups, and to propose what transparency requirements should be applied to each of those levels. We believe that, generally speaking, the less private a space, the greater the requirement for transparency and access.

At the outset, it is imperative to note that we urge further consultation on this issue between the Government and relevant legal, civil society and research experts beyond the official consultation period for the White Paper before any definitions of public or private online spaces are taken forward. Both in definitional and legal terms, the public vs private debate is extremely complex, and we urge the government to commit to significant further discussion, and hope this submission acts as a provocation to this.

How do we determine public vs. private?

The average internet user will use a range of digital services that run the gamut from public to private, including multiple services on a single platform. Posting a Tweet, for instance, would be regarded as a more public action than sending a private message on the same platform. It is right for the white paper to reflect the differences between these functions. However, the distinction between the two and the grey area are not easily defined.

We understand there to be at least three levels of expectations of privacy represented by the functions of online communication channels. A binary conception of private and public online is too limited to represent the reality of communications online. These three levels can be linked to subsequent expectations of privacy and transparency, which are explored later in this document.

The factors determining whether a space should be publicly observable are many, and a regulator should include and embrace nuance and complexity in any processes set-up to determine whether a space should be considered public or private. To determine a communication channel's level of public/private obligations, the following set of criteria should be examined as part of a decision.

When used in conjunction, these tests could help to evaluate the level of privacy of an online service or communication channel. This list of tests is not comprehensive: further test criteria should be explored to confidently and fully evaluate a service's privacy status.

1. Size of channel

The size of a channel may be a useful indicator of whether its members feel it is a private space. For instance, a Facebook group with thousands of members sharing content is less likely to be forum in which privacy can be reasonably expected. Tests could look to measure:

- The number of users involved in sharing content.

- The potential reach of users sharing content to other members.

2. Purpose/rules of channel

It is likely that the stated purpose of a channel is a good indicator of whether or not its users believe it to be a private space. For instance, a page described as a space for sharing national news and provoking public debate is more likely to be a public space than a page for a local neighbourhood watch group.

3. Accessibility of the communication channel

The presence or level of barriers to public participation in an online space may also act as a useful indicator for whether users might expect the space to be private. These barriers may include:

- Password protection
- 'Invite only' / the necessity to possess a URL link to enter the online space
- Registration / requirement for an account on the service
- Indexable communications, i.e. whether communications on the service are accessible for the public through public indexes such as search engines

4. Nature of relationships within the communications channel

A page or group may bring together people who know each other in other ways, or may connect strangers. Content shared may also be directed towards 'friends' on a social network or the network as a whole, likely impacting how public or private that content is perceived by a user.

- Existing relationships/contacts vs. 'strangers', i.e. nature of relationships within service
- Mechanics and decisions around breadth of sharing

Types of space

Fully public services: No reasonable barriers to research

We believe there are spaces online that platforms should recognise as public, and with governmental direction work to ensure those spaces are computationally observable by its users, including academia and civil society organisations. Content and communications posted to these spaces should be treated as being contributions to a public debate. We believe these public spaces tend to be indexable by search engines, with little or no access requirement for accessing or being exposed to that content. Examples of these kinds of spaces would include Wikipedia, JustPaste.It, the main Twitter timeline, a 'public' Facebook page, or comments below-the-line on free-to-access news sites.

Fully private services: Outside the remit of research

In a recent legal case, BC and others v Chief Constable Police Service of Scotland, officers' use of a WhatsApp group revealed potential misconduct. Internal misconduct charges made by the Professional Standards Department were then challenged by the accused, arguing that "since the messages were sent in a private group, their use in the context of misconduct proceedings against them amounted to a violation of their privacy".[15] In this case, the court accepted the group was a private space (but to note, did not come to a decision as to its use in misconduct proceedings).

Conversely, there are private functions which under law must remain outside the scope for potential research and public visibility. These communications are protected by Article 8 of the ECHR and in UK law. As such, these spaces tend carry significantly higher expectations of privacy on the part of their users and significantly higher barriers to observation. It is our understanding that observation of these communications requires a warrant, investigations under the Investigatory Powers Act 2016, and are not within the remit of a regulator. One-to-one communications that have a barrier to accessibility for additional users have long been included within constitutional and international law-based guarantees for privacy. We do not believe that it is the place of this regulator to monitor these spaces.

However, in answer to question ***7a. What specific requirements might it be appropriate to apply to private channels and forums in order to tackle online harms?***, we recommend moving away from a focus on content and towards encouraging better platform design. The decision by WhatsApp, for instance, to limit the number of times a message can be forwarded to five times is a useful example of how a change in a platform's architecture can improve its function without requiring monitoring of communications. These changes should be evidence-based and evaluated for impact as far as possible. We recommend commissioning further research into the ways in which platforms can build architectures that reflect best practice.

Examples of these kinds of spaces would include one-to-one WhatsApp or iMessage communications, emails, VoIP telephone conversations.

Other online communication services: Unknown or unclear accessibility to research

There is a large grey area between those services we believe to be either demonstrably public or private. These include limited-access or closed groups with multiple participants. In all grey zone cases, a service should be first assumed to be private, and any potential status as a 'public' service must be proved.

Examples of these spaces include Discord channels, Facebook 'closed' groups, Slack Channels, WhatsApp groups, and Google documents.

[15] https://www.scotcourts.gov.uk/docs/default-source/cos-general-docs/pdf-docs-for-opinions/2018csoh104.pdf?sfvrsn=0

Conclusions

There is no simple solution to this question, and further consultation is vital. We urge DCMS to champion efforts by civil society and research organisations to research and report on what is happening online, and to commit to working with these groups and the industry to agree how this research can best be supported.

Safety by Design

The regulator should provide guidance in a range of areas to ensure effective safety by design, including but not limited to:

- AI and algorithmic decision-making
- User journeys and uses of positive "nudges" that encourage user choice and clear understandings of settings (e.g. privacy)
- Recommendation filters
- Live-streaming

In order to achieve safety by design in these areas, and others, organisations should be encouraged to employ scenario or risk forecasting and stress-testing approaches to ensure products are future proofed against online harms. Practical guidance should also be provided on the key principles of ethical design, as well as practical tools to help assess and predict possible safety concerns during the early stages of design. Overall, the regulator should encourage a cultural shift in the online technology sector to ensure that safety by design becomes embedded as a foundational principle throughout technology companies.

Educating and Empowering Users

Overall the government should play a key role in education and public awareness activities, supported through revenues raised via a new digital tax on online services. Overall, the public requires simple, common and consistent language and indicators across the online environment to help inform their choices and improve their understanding of the products they use, the spaces in which they participate, and their rights, responsibilities, and options for redress. This could include the development of a 'highway code' type system for the online environment, including commonly understood symbols designating the types of spaces found online, and the relevant rules and regulations that apply, supported by public awareness campaigns and clear guidance, and potentially online training modules. For example, Creative Commons have been relatively successful in designing a similar system for certain types of intellectual property online.

Additionally, those with the influence to support users in becoming empowered digital citizens have a responsibility to do so. Technology companies, governments, educators, parents and civil society actors need to work together in order to keep pace with the changes

to the digital world and update the education system accordingly. While there is broad recognition of the need to build digital literacy skills and knowledge, as evidenced in the White Paper, stakeholders must go beyond developing digital literacy and focus on the norms and behaviour that comprise *digital citizenship*.

The following recommendations focus on how further collaboration between stakeholders can empower young people to act on the agency they have to improve their online communities as good digital citizens:

- **The UK Government should define and standardise digital citizenship** to enable educators to understand what it is and recognise its importance. The Government's recent proposals for a media literacy strategy should sit at the heart of a wider drive to increase digital citizenship learning.
- **Digital citizenship should be embedded into the national curriculum**, with more specific guidance and training for practitioners on how best to teach it, and through which programme of study it would most effectively be taught. Government should encourage and support school and youth centre leaders to train their staff to deliver digital citizenship learning effectively, combining this training within initial teacher training, continuous professional development and youth worker training.
- **All stakeholders in digital education should co-ordinate more effectively** to ensure teaching and learning keeps pace with changes in technology and reflects the nature of contemporary online harms.
- **Digital citizenship education models should be tailored for delivery in informal education contexts**, where in-depth conversations and inspiring practitioners can effect positive behavioural change online.
- **The UK Government should give schools adequate guidance** on how to spread digital citizenship across the key stages, to ensure that gaps do not emerge in students' learning and that knowledge, skills, behaviour and attitudes are developed each year.

The proposed regulator may play a role in informing these efforts, based on the information and insight available to it, but the onus should be on Government departments such as the DFE and DCMS.

Regulatory Scope

Our response approaches questions of regulatory scope from two perspectives: whether the scope of the proposed regulatory framework is effective and proportionate in principle; and whether the scope of the regulatory framework would be effective and proportionate in practice.

In principle:

Scope: Harms and illegal activities are conducted through an extremely broad spectrum of technology platforms and services, as evidenced in ISD's extensive research on disinformation and extremist or terrorist use of the internet.[16] The wide scope of platforms that would be implicated in the duty of care is, in principle, a necessity to comprehensively and sustainably address the evolving tactics of purveyors of online harm, who do not act solely on the few, largest technology platforms, but instead use an entire ecosystem of platforms to conduct harmful activity. A focus on just a few large platforms would be limited: the focus of improving content moderation approaches on a few large platforms over the past three years has led to a platform migration of many purveyors of hate speech, extremism, terrorist content and disinformation away from large platforms to smaller platforms with little or no oversight, limited or no Terms of Service (e.g. Gab), or in some cases, any appetite or intent to respond to online harms (e.g. 8chan). A limited focus on the few largest platforms would simply accelerate this phenomenon.

Proportional approach: Given the breadth of platforms and services included in the scope of the framework, the government should commit to a genuine interpretation of the duty of care approach provided by Carnegie in their consultations to the government. This should focus on a regulatory framework of risk management and harm mitigation rather than direct liability to ensure that it is a proportionate approach for the range of platforms and services included in the regulator's remit. The government should focus efforts on an approach to manage potential harms while ensuring that the level of responsibility and activity required of platforms/services remains achievable for platforms/services of a wide range of sizes and resource/capacity, so as not to strangle the potential innovation and growth of the UK tech sector and to avoid discouraging technology platforms from providing access to UK users. ISD emphasises the need to focus on a 'safety by design' approach to regulation, only briefly fleshed out in the White Paper.

16

- https://www.isdglobal.org/isd-publications/briefing-note-el-rubio-lives-the-challenge-of-arabic-language-extremist-content-on-social-media-platforms/
- https://www.isdglobal.org/isd-publications/disrupted-evidence-of-widespread-digital-disruption-of-the-2019-european-elections-joint-submission-by-avaaz-and-isd/
- https://www.isdglobal.org/isd-publications/interim-briefing-propaganda-and-digital-campaigning-in-the-eu-elections/
- https://www.isdglobal.org/isd-publications/battle-for-bavaria/
- https://www.isdglobal.org/isd-publications/smearing-sweden-international-influence-campaigns-in-the-2018-swedish-election/
- https://www.isdglobal.org/isd-publications/make-germany-great-again-kremlin-alt-right-and-international-influences-in-the-2017-german-elections/
- https://www.isdglobal.org/isd-publications/the-fringe-insurgency-connectivity-convergence-and-mainstreaming-of-the-extreme-right/

In practice:

Scope: Given the scale of user-based content uploaded online every minute and the range of platforms used to host such content, there is no doubt that the regulator will struggle to consistently apply and enforce regulatory standards. A system of triage or prioritisation will be needed to help identify the most pertinent threats to individuals or to society, which might include assessment of the number of users per platform (and potentially the time spent by users on the platform). However, this measure of prioritisation should not rely solely on assessments of user-bases, as some relatively small platforms have been shown to play a crucial role in hosting and sharing content supporting or planning extreme acts of harm such as terrorist attacks, for example in the context of the Christchurch terrorist attack announced and supported on the 8chan platform. While this issue of platform scope might apply especially seriously to consistent enforcement of content moderation regulation, issues of platform scope will also hinder a consistent application of 'safety by design' regulation.

Enforcement for small platforms or systems: Amid the range of platforms hosting illegal and 'legal but harmful' content aimed at UK audiences are a number of purpose-built systems that aim to avoid any form of oversight or content moderation or to comply with national law. These include a number of platforms and systems built to host and share extremist and terrorist material. Many of these platforms have opaque ownership structures. The regulator will need to consider how enforcement obstacles that such platforms attempt to put in place can be overcome without compromising legitimate freedom of speech, deterring innovation via barriers to entry, or disproportionately disincentivising the provision of services to UK users.

Regulatory Proportionality

Firstly, ISD cautions against application of the entire scope of potential regulatory action to the overly broad spectrum of 'legal harms' included in the White Paper. 'Legal harms' should be treated distinctly from regulation of 'illegal harms' in order to protect well-established rights to speech and privacy. It is now well-evidenced that some 'legal harms', such as disinformation and extremism (that does not meet the threshold of hate speech), pose threats to public safety, public health and the integrity of democratic processes. However, in seeking to deal with these 'legal harms', ISD urges the strong application of transparency and safety by design regulation over the *systems* and *processes* of technology platforms, rather than a focus on content moderation regulation, which would endanger rights to free speech if applied to this set of 'legal harms'. The content moderation regulation suggested as part of the regulator's activities should apply to 'illegal harms' only.

Additionally, ISD also urges the proposed new regulator:

- To work closely in cooperation with existing regulators with relevant experience and related remits, such as the Electoral Commission and the ICO.

- To ensure in-house expertise and experience is drawn from a wide range of fields and sectors, to avoid creating a 'revolving door' between the regulator and industry to avoid the risks of regulatory capture.
- To maintain strong relationships with experts, researchers and civil society organisations to ensure responding to accurate understandings of threats to rights and safety.
- To ensure thorough transparency reporting by the regulator itself, including information on its processes, methods, decisions, and outcomes.

Regulatory Oversight

Parliament should have the key role scrutinising the regulator and its work, including the codes of practice discussed in the White Paper. All elements of the regulator's work should fall within the scope of parliamentary oversight, regardless of the issue area concerned. None of the codes of practice should be developed outside of parliamentary oversight and scrutiny, including the codes of practice relating to terrorist content and CSE content, as suggested in the White Paper.

If government regulation of online harms is truly designed with the purpose of protecting democracy and human rights, full oversight of the regulator through a democratically elected parliament is vital. ISD, as a leading counter-extremism NGO that works closely with both government and the private sector, emphasises the responsibility of platforms to act appropriately and proportionately on illegal and harmful content. Similarly, the objectives of the government to fight serious online harms, such as terrorism and CSE, are crucial.

However, in some cases they can lead to disproportionate measures, as the recent example of Archive.org shows. In April 2019, the website was urged by the EU Internet Referral Unit to take down 550 URLs, including content from C-SPAN and academic articles, all mistakenly identified as terrorist propaganda. This exemplifies how unchecked counter-measurements against serious (and illegal) harms - such as terrorism - incorrectly implemented by the governments can be problematic.

The implementation of sufficient checks and balances in the form of parliamentary oversight of the new regulatory framework is therefore crucial. Acknowledging the wide-ranging scope and impact of the proposed regulation, unchecked implementation of disproportionate measures as shown above might become a serious threat for fundamental human rights such as freedom of expression. The suggestion of direct control by the Home Secretary over certain codes of practice directly contradict the democratic underpinnings of the regulatory objectives and lack sufficient transparency and oversight.