Research into Online Platforms' Operating Models and Management of Online Harms

June 2019

ICF makes big things possible

ICF is a global consulting and technology services provider with more than 5,000 professionals focused on making big things possible for our clients. We are policy specialists, social scientists, business analysts, technologists, researchers, digital strategists and creatives. Since 1969 government and commercial clients have worked with ICF to overcome their toughest challenges on issues that matter profoundly to their success. Our five core service areas are described below. Engage with us at icf.com.

Research + Analyse

Our teams delve deep into critical policy, industry and stakeholder issues, trends, and behaviour. By collecting and analysing data of all kinds, we help clients understand the current landscape clearly and plan their next steps wisely.

Assess + Advise

With equal parts experience and dedication, our experts get to the heart of the issue—asking all the right questions from the start. After examining the results and evaluating the impact of research findings, we counsel clients on how to best navigate societal, market, business, communications, and technology challenges.

Design + Manage

We design, develop and manage plans, frameworks, programmes, and tools that are key to each client's mission or business performance. These solutions often stem from our analytics and advice.

Identify+Implement

Our experts define and put into place the technology systems and business tools that make our clients' enterprises more effective and efficient. We deploy standard or customised methodologies based on the business context.

Engage

Realising the promise of the digital revolution requires foresight and heightened understanding. Both are baked into the solutions-focused engagement work that runs through all we do.

Research into Online Platforms' Operating Models and Management of Online Harms

A report submitted by ICF Consulting Services Limited

June 2019

ICF Consulting Services Limited
Watling House
33 Cannon Street
London
EC4M 5SB

T +44 (0)20 3096 4800
F +44 (0)20 3368 6960

www.icf.com

Acknowledgement

This study was conducted by ICF Consulting Services Ltd on behalf of the UK Department for Digital, Culture, Media and Sport (DCMS). The core research team that delivered the study consisted of Dr Simona Milio, Kate Regan, Facundo Herrera, Luca Favero and Helen Quaresma. The study benefited from the specialist input of its expert advisory board: Dr Victoria Baines, Professor Ian Walden and Michael Moran.

The authors would particularly like to thank those platforms that participated for the time and support they offered throughout the study.

Contents

1 Introduction

1.1 Aims and scope of the study

ICF was contracted by the UK Department for Digital, Culture, Media and Sport (DCMS) to undertake research into how online platforms operate to tackle online harms, determining their incentives and capabilities for doing so and how they could adapt to potential regulation.

The overall aim of this research was to provide a review of online platforms used by UK citizens to understand the key operating and financial information of different platforms and how these might affect approaches to prevent online harms. The key research objectives were:

- To understand what different platforms define as harm on their platforms;
- To understand the incentives, capabilities and methods of different platforms to address online harms, including technical and economic capabilities;
- To measure how effective these incentives and capabilities have been at reducing harm;
- To understand the potential impact of regulation on these platforms.

An Expert Advisory Board (EAB) was recruited to provide support at strategic moments throughout the study. Information on the EAB composition can be found in Annex 5.

Ensuring the anonymity of platforms in our research

In this report, online platforms are referred to by number: online platform 1; online platform 2; and so on. This is to ensure the anonymity of the platforms.

Online platform 5 did not agree to the publication of any information they shared during the interview. Consequently, this report refers to platforms collectively as being 11 up until Section 2.1.3, as this information is drawn from publicly available information.

From Section 2.1.3 onwards, the report refers to the platforms collectively as 10, to reflect the request to withdraw qualitative information that was given by online platform 5 during the study, and any reference to the platform has been removed.

1.2 Purpose and structure of the report

This purpose of this report is to present the final outcomes of the research tasks undertaken throughout the study. The main report follows the following structure:

- Section 2. Defining online harms.
  - Overview of the **types of harm categories which surfaced during the literature review,** and which are relevant to the study.
  - Presentation of **typology of harms** and development of **taxonomy of harms** which enabled harms to be categorised across the platforms.
  - The extent to which platforms differentiate between **lawful and illegal harms**.
  - Description of the ways **platforms amend their own internal categorisation of harms.**

- **Section 3. Understanding the strategies, capabilities and incentives of platforms to address online harms.**
  - Presentation of the **strategies and methods employed to address harms**, capturing: general approaches to safety on the platform; user reporting mechanisms in place; the technological tools in place to tackle harms; and where data is available, the economic and human capabilities of platforms to tackle harms.
  - A description of how the 11 platforms reflected that they are **currently incentivised** to tackle online harm, and how they might be incentivised to do so further.

- **Section 4. Understanding effectiveness and efficiency in how platforms tackle online harms.**
  - Description of how platforms **perceive their own effectiveness and efficiency**.
  - Description of **metrics** used by platforms to measure effectiveness and efficiency, where survey data is available.
  - Description of whether platform respondents perceived there to be **a relationship between incentives, capabilities and effectiveness.**

- **Section 5. Impact of regulation and other factors on reducing online harms.**
  - **International regulatory review** (of Germany, Australia and France) to determine whether the regulatory approaches employed in those countries have affected how online harms are tackled by online platforms.
  - Presentation of **the perceived economic impacts of address online harms.**
  - Presentation of **the perceived impact** of both **technological tools** and **transparency reporting** on platforms.

- **Section 6. Conclusions and business model.**
  - Presentation of conclusions drawing from the research and reflecting the key research objectives.
  - Explanation of the business model, developed to establish a conceptual overview of the main business processes relevant to how an online platform operates to tackle online harms.

The following Annexes are included:

- Annex 1 Literature review
- Annex 2 Research framework
- Annex 3 Limitations and mitigating measures
- Annex 4 Reference list
- Annex 5 Expert Advisory Board

1.3 Research framework and methodology

1.3.1 Research framework

The research framework was developed by mapping 14 main research questions and 20 sub questions onto the four key research objectives presented in the section above. This process was guided by the Expert Advisory Board and a result of the scoping interviews undertaken during the Inception Phase. The research questions have been answered throughout the different phases of the study and are presented fully in Annex 2.

The methodology was designed so that tasks were divided into three main stages: the **Inception Phase,** the **Interim Phase**, and the **Reporting Phase**. Those first two phases, and the tasks they involved, are detailed below.

1.3.2 Inception Phase

The first task of the Inception Phase was to **agree on a sample of 11 platforms** for inclusion in the study, based on an original selection of 25. The first part of a **two-fold preliminary data and literature review** was then undertaken. This consisted of both a compilation of the broader contextual literature and theory on the various elements involved in reducing online harms, and a collation of Terms of Service, community policy documents[1], and transparency reports for each of the 11 platforms. Finally, three **scoping interviews** were undertaken with stakeholders in the area. These included two representatives of online platforms (one gaming; one social networking) and one with an online safety expert. The purpose of the interviews was to enhance the knowledge base of the research team before designing and developing research tools, ensuring that they captured relevant contextual factors and specific processes intrinsic to addressing online harms. Several issues raised during the scoping interviews were later reflected in the full interviews undertaken with platform representatives and so helped inform and refine the topic guides and questionnaire.

1.3.3 Implementation Phase

The Implementation Phase consisted of the second part of the **two-fold data and literature review**: the analysis of the identified literature in alignment with the research framework, and the analysis of the platform operating documents previously identified.

An **interview** was conducted with at least one representative of each platform. The purpose of the interview was to gather information related to the platform's perceived effectiveness and efficiency in reducing online harms, current and potential incentives for reducing harm, resource allocation and costs involved in reducing harms, and the impact of internal and external measures in doing so. The interviews meant more qualitative and sensitive information could be elicited.

Based on the information received during the interview, the **targeted platform survey** was amended to reflect the harms prioritised by each platform uniquely. The survey captured the moderation processes employed by each platform to tackle certain online harms. The survey was structured *per harm*; for each harm, the platform representative was asked to elaborate on the different moderation strategy/ strategies that allow platforms to address that harm. The harms which were included in the survey were determined based on whether they are in scope of the study and confirmed at interview stage[2].

[1] The term community policy document is used to capture documents which outlines the policies, rules and guidelines that platforms users are expected to adhere to when using the platform.

[2] At the time of submission of this report, only 4 platforms had completed the survey via the survey platform: online platforms 1, 4, 9 and 10. Online platform 7 preferred to submit a narrative response to the survey in order to 'explain more fully its wider approach to managing content on its platform and present its use of human review and moderation technology in this context'. Consequently, due to the limited responses received, conclusions are not drawn based on survey data, although information provided through the survey was still valuable and is explored in Sections 3 & 4. Other reasons for non-completion of the survey included that: one platform did not feel that the survey questionnaire reflected frameworks they used to tackle harm; two platforms stated that they had provided the content of the survey information already provided to DCMS. The other platforms did not provide a reason for not completing the survey.

The outcomes of the tasks undertaken enabled the processes, capabilities and incentives which drive how each platform included in the study tackles online harms to be described and mapped.

1.4 Limitations and mitigation measures

Whenever a limitation was encountered throughout the study, a measure was employed to mitigate its impact. While the collaboration with the platforms was constructive throughout the study, the sensitive nature of certain topic areas meant that limitations were encountered. A main limitation related to the lack of available data about highly sensitive financial information. Likewise, data on staff allocation to moderation activities was limited without any financial value attached: this made it impossible to estimate the cost of tackling online harms. In addition, respondents were reluctant to share evidence on their business strategies, customer segmentation or unique selling points that would make the online platform identifiable. Finally, the content covered during consultation with platforms was self-reported and so risks a certain degree of bias.

A range of mitigation measures were employed to address those limitations. A table which fully describes each limitation and the accompanying mitigation measure that was undertaken are presented in Annex 2.

2 Defining online harms

Key messages

- Across the 11 platforms included in the research, harms (those within study scope) can be grouped into 12 *thematic harm* categories and 10 *associated* harms.
- The thematic harms which are replicated the most across platforms relate to adult sexual content, violent content and conduct, and endangerment of children. These are harms for which content is often image-based, or which relate to activity which is nearly always illegal, irrespective of context.
- The thematic harms which are replicated the least frequently across platforms are threat, criminal content and conduct (as a stand-alone harm), impersonation and fake news and representation. Apart from criminal content and conduct (see below for more in-depth analysis of this harm category), these are harms whose severity and legality require a greater analysis of contextual factors.
- Platforms do not tend to differentiate between illegal and lawful harms in their operating documents and policies.
- Platforms amend how they categorise and define harms by involving a range of internal actors.

2.1.1 Introduction

The purpose of Section 2 is to understand how platforms define harm, identifying where there are differences in definition, whether platforms officially differentiate between illegal and lawful content, and to determine whether platforms have formal mechanisms in place to amend how harms are categorised internally.

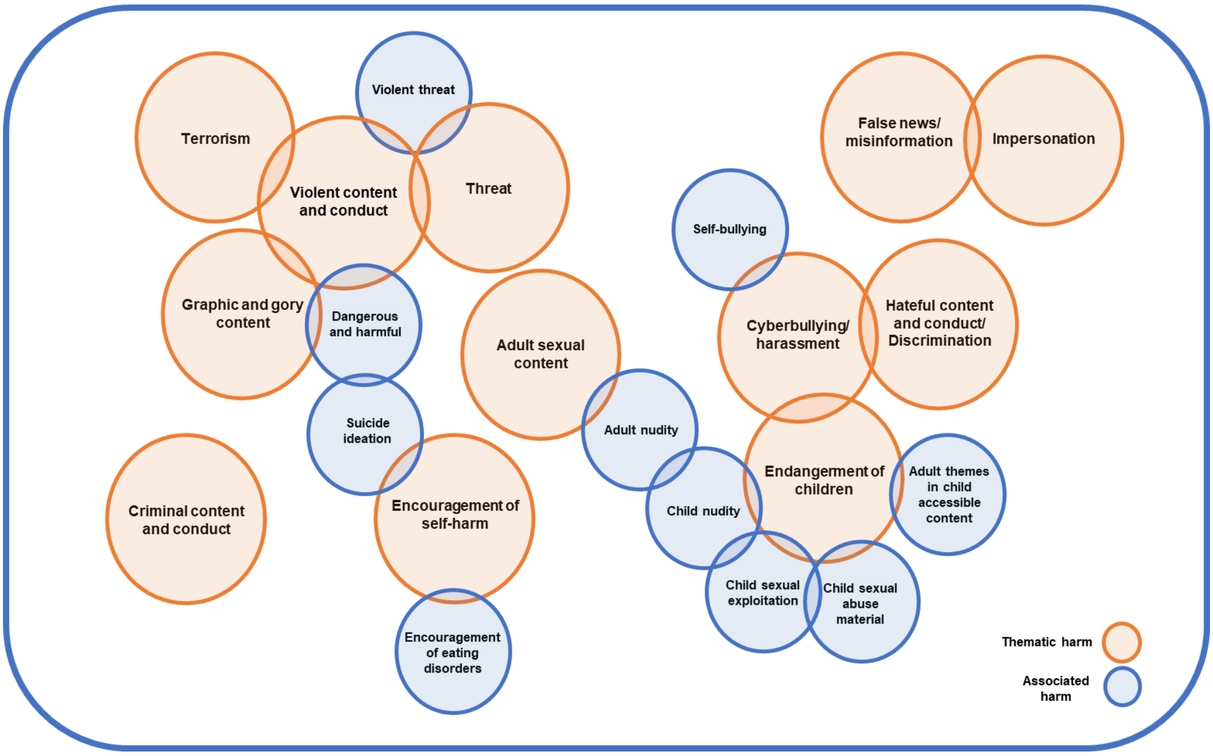
2.1.2 Definition and categorisation of harms across platforms

In early inception discussions with DCMS, a typology of harms was shared to help establish which harms would be relevant to the study.

A review was undertaken of all harms highlighted in the policies across the Terms of Service and Use documents and community policy documents of the 11 platforms, to gather a full list of those within the scope of the working **typology** of harms. Across the 11 platforms, 92 individual harms were identified, of which there was considerable overlap in terms of their meaning and scope. A **taxonomy** was created to form categories which adequately captured those 92 harms. By collating harms by type, 12 common *thematic harms* were identified which featured most widely across all 11 platforms. These were: terrorism; violent content and conduct; graphic and gory content; threat; adult sexual content; encouragement of self-harm; criminal content and conduct; cyberbullying/ harassment; endangerment of children; hateful content and conduct/ discrimination; false news/ misinformation; and impersonation. At a more granular level, 10 *associated harms* (which are more distinct) were identified within these broader categories. These were: adult nudity; child nudity; child sexual abuse material; child sexual exploitation; adult themes in child accessible content; self-bullying; suicide ideation; dangerous and harmful content; and violent threat.

Figure 2.1 presents, in a Venn diagram illustration, the conceptual overlap between thematic harms and associated harms. The categorisation of harms (by thematic harm and associated harm) was an important way to interpret survey data without revealing exact wording and phrases used by different platforms to define harm, which would violate platform anonymity. For this same reason, the original harm taxonomy cannot be reproduced here.

Figure 2.1 Categorisation of policies by harm across platforms.


Terrorism
Violent threat
Violent content and conduct
Threat
Graphic and gory content
Dangerous and harmful
Adult sexual content
Suicide ideation
Encouragement of self-harm
Criminal content and conduct
Encouragement of eating disorders
False news/ misinformation
Impersonation
Self-bullying
Cyberbullying/ harassment
Hateful content and conduct/ Discrimination
Adult nudity
Child nudity
Endangerment of children
Adult themes in child accessible content
Child sexual exploitation
Child sexual abuse material
Thematic harm
Associated harm

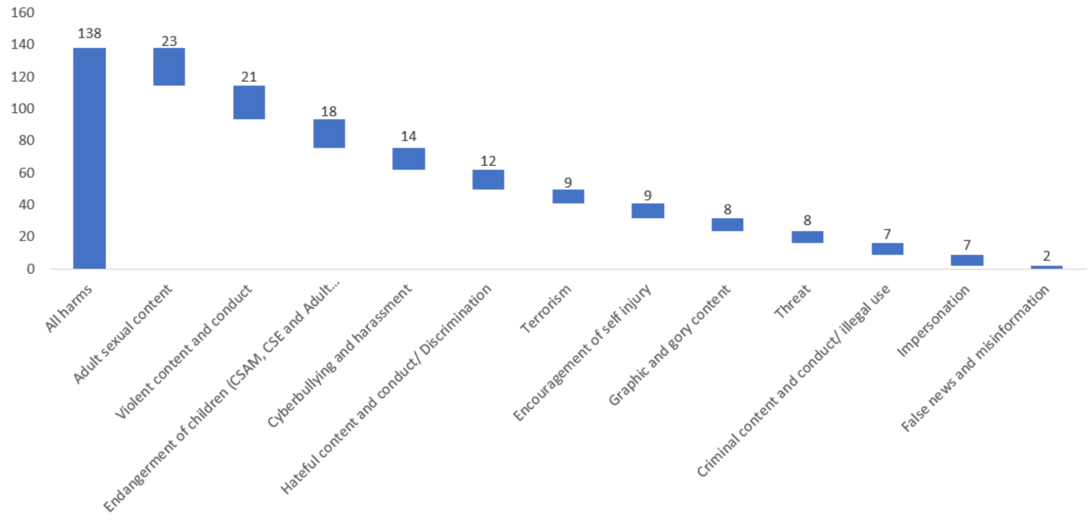
Violent content and conduct is commonly identified across platforms as harmful, and collectively has significant reach in terms of its relevance to other harms. Its associated harms include violent threat, dangerous and harmful content. It is also linked with other thematic harms, including terrorism, graphic and gory content and threat. Endangerment of children, as the Venn diagram shows, is a harm that captures a range of varied harms which affect children: child nudity; child sexual exploitation; child sexual abuse material and adult themes in child accessible content.

Criminal content and conduct is presented here as a stand-alone thematic harm. Across the 11 platforms, seven referred to illegal or criminal content or conduct: five platforms defined illegal or criminal content or conduct as a catch-all stand-alone harm, while other platforms referred to criminality in the context of a known illegal harm (trafficking) or more broadly, in relation to dangerous content. As most main harms and associated harms presented in the diagram could constitute illegal content or conduct depending on the particular situation, the potential associations between harms has not been drawn.

Figure 2.2 depicts how frequently the 92 harms (identified across the 11 platforms) were marked as relevant to each thematic harm across all relevant operating documents. In some cases, a harm, by its particular definition, would be marked as relevant against more than one thematic harm; consequently, the total number of harm ‘entries’ totalled 138, not 92[3].

[3] For example, ‘terrorism’ was mentioned across platforms as a stand-alone harm, in relation to violent content, and in relation to criminal content.

Figure 2.2 Chart depicting the most commonly featured thematic harms



As Figure 2.2 shows, the most commonly featured harm is adult and sexual content (with 23 entries), followed by violent content and conduct (21 entries), then endangerment of children (18 entries). This is undoubtedly because of the scope of these main harm categories; adult sexual content is relevant to harms with a focus on adult nudity, sexual content, sexual solicitation and sexual exploitation of adults, amongst others. Violent content and conduct includes violent and disturbing images, the promotion of violence, violence of the grounds of a protected characteristic and violent threat, amongst others. Endangerment of children included children accessing adult content, child sexual exploitation and child nudity, amongst others. False news and misinformation was by far the least commonly featured harm across platforms, followed by impersonation and criminal content and threat. It should be noted that criminal content and threat, as a main harm, had the potential to register a far greater number of the 92 harms identified across the platforms; however, this would of course depend on the situation and would thus be hypothetical. Consequently, for this category, only harms which specifically mentioned illegal or criminal behaviour were counted.

Illegal and lawful differentiation in operational documents

Across the online platforms interviewed during the study, all have a public facing document or webpage aimed at their user community in which platform policies and rules are defined and elaborated, in addition to a Terms of Service or Use document. Of the Terms of Service documents of the 11 platforms, nine directly state that compliance with the community policy document must be agreed to as a condition of using the platform. Online platforms 2 and 4 incorporate the rules outlined in the community policy document into the Terms of Service, rather than citing it as a separate document.

Beyond the (public law) obligations outlined in Terms of Service or Use documents which the platform user and the platform itself are bound by, the community policy documents of all the platforms provide a more user-friendly presentation of prohibited behaviour or content. Only one online platform provides an uncategorised list of prohibited content, while all others group prohibited behaviour and risk by categories of harm. Across the community policy documents, six platforms explicitly ask their users to refrain from posting illegal content or engaging in illegal activity on the platform; however, there is no further elaboration beyond this rule, and so serves as a catch-all term for any illegal activity. Regardless of including a rule prohibiting illegal behaviour and content, those six platforms, in addition to the other five, categorise harms in a way in which potentially illegal and lawful harms can be grouped together, depending on the case in point; for example: harassment and cyberbullying, sexual content, and graphic content are all typical categories used across platforms in their community policy documents. Within these categories, content or behaviour may be illegal or lawful, but distinctions are not made in the categorisation themselves; general language further allows platforms to allow for jurisdictional differences across the countries they operate in. Thus, current categorisation in place by platforms allow both illegal and lawful harms to be captured.

2.1.3 Amending the categorisation of harms: the platforms perspective

Platforms reflected the need to be adaptable to new challenges and trends as they emerged; online platforms 3 and 4 cited the ‘Tide Pod challenge’[4] as one such unanticipated phenomenon which posed a risk to its users. Both platforms stressed the importance of having policies in place across Terms of Service documents that would be able to capture such unpredictable risks.

Online platform 1 highlighted that moderators were the ‘first line of defence’ when it came to identifying new harms, whether undertaken by the in-house moderators responsible for language content, or third-party moderators responsible for moderating photos and image-based content, suggesting a good working relationship and correspondence with both internal and external staff. Online platform 6 also expressed that moderators were the key actors in analysing UGC trends to influence how harms manifested on the platform, and accordingly how harmful phenomena was understood and categorised on the platform.

Six platforms explicitly mentioned that the process of amending how harms were defined and categorised involved both internal and external factors and actors. Online platform 2 mentioned that they could amend harm categories and polices by combining the analysis of trends undertaken by moderators with feedback received from an extensive communication strategy the platform has with its users. This communication strategy consists of a weekly streaming service between users and the platform in which the users are asked to report on issues or ideas they had for the platform, while also encouraging users to send messages (via email or online form) to raise any issues they face on the platform. In addition to this, the platform monitors media coverage of emerging and existing harms. Online platform 7 highlighted that they monitor other companies to understand how they define their guidelines and handle online harms. This is one of many factors which inform how they develop their own polices.

Online platform 11 elaborated on the extensive process in place by which policies around harms are refined or added to. They highlighted that once a new challenge was presented (by an NGO partner, law enforcement agency, government or media), a working group process led by content policy generalists but with representatives from across the company will draw from feedback and data projections to propose recommendations to the policy in question. Based on these a decision would be made on whether to enforce a change to the policy.

A representative of the same online platform reflected on the fact that the harms they define in their Community Guidelines are a means to brand their polices; in the external world, they felt that conceiving of and tackling harms could not be done in such a systematic manner. They added that harmful behaviour online is not exclusively an online phenomenon, but that the user responsible for generating harmful content or behaviour is using an online tool to do something that they may have nevertheless undertaken offline. They further elaborated that there are certain harms that it is wholly their responsibility to tackle: spam or misrepresentation. Other harms necessarily involved the involvement of other actors. False news was cited as one such harm; they felt that deleting all false new content wasn’t a possibility, and that consensus had to be reached on whether it was false or not in the first instance. A representative from online platform 7 highlighted a similar point, stating that responding to harmful content online was an

[4] A viral trend in which internet users would film themselves eating a detergent capsule (usually of the ‘Tide’ brand).

‘ecosystem challenge’ that relied on action by access providers, hardware manufacturers, software developers in addition to the platforms themselves, and which must also be reflected in government policy.

3 Understanding the strategies, capabilities and incentives of platforms to address online harms

3.1 Strategies and methods of platforms to address online harms

Key messages

- Platforms enhance their **general approaches to safety** (beyond just moderating content) in a range of ways which include: employing the expertise of internal and external specialists and providing tools for parents and teachers.
- A platform's recognition of its responsibility to its users can both *strengthen* **approaches to safety**, or make them more *limited*, depending on the user base and the unique value of the platform.
- In line with expectations, all platforms that responded to the survey employ a combination of **pre-moderation, post moderation and reactive report moderation** strategies to moderate harmful content
- All survey responding platforms reported having procedures in place for **law enforcement authorities to report** content which is locally allegedly illegal.
- **Reporting violating** users to local authorities is common practice among platforms, though only for certain harms. Not all platforms have done so in the last year.
- Platforms typically follow a standard procedure regarding **user reporting**. Some platforms vary in the extent to which they allow users to interact with the platform or explain reported content in more detail.
- Survey data highlights that for all responding platforms, users found to be culpable of activities or behaviours which constitute any of the harms they moderate might lead to **account suspension.**

3.1.1 Introduction

The purpose of Section 3.1 is to explore the general approaches to online safety that platforms reported to observe, expanding on the reasons behind them. It also aims to identify the moderation strategies platforms use to tackle harm, the types of automated processes involved in moderation, whether user reporting functions differ across platforms, whether platforms report users to local authorities and whether users are blocked; and the conditions under which a user's account is suspended.

General approaches to safety across platforms

The online platforms consulted through interviews tended to highlight an approach to ensuring safety that went beyond solely filtering or removing illegal, policy violating or undesirable content. Those approaches involved:

- employing the expertise of internal and external specialists in the area;
- providing tools for parents and teachers;
- recognising the responsibility of platforms to their varying user-bases, whilst also acknowledging how the different countries the platform operates in will have specific norms to be respectful of.

Employing experts

Four platforms out of 10 highlighted during interviews that they had an approach to user safeguarding that involved drawing on the professional experiences of internal platform personnel, or by collaborating with external experts to enhance their efforts in this area.

A representative of online platform 1 with responsibilities for child safety indicated during interview how their own experience as a social worker and working at charities with a focus on child rights, strengthened the platform's approach to child online safeguarding, especially when this expertise was paired with the experience and capabilities of engineers, designers, customer care and support team staff, and moderators. Online platform 4, which has a young user base, cited that they have funded research into online phenomena and have additionally enlisted the support of academic professionals to support them in their efforts to ensure child safety on their platform. Online platform 7 reflected on how their efforts in ensuring protection from child sexual exploitation had particularly grown recently, with former employees of the US non-profit organisation National Centre for Missing and Exploited Children working on child safety there. In addition, that same online platform reported that it receives and shares threat information to peer companies, external consultants and where appropriate law enforcement.

Platform 7 mentioned that it consults with experts and civil society organisations, such as the Samaritans to ensure the platform receives input from various communities/ perspectives in designing policy and processes. These partnerships also help ensure that their community of users is engaged and knowledgeable of certain harms and was seen to be a particularly important strategy considering the inherent limitations of technology in detecting most harms. User engagement is further enhanced by the platform's promotion of its own campaigns around a particular issue, or by highlighting certain International Days. Community engagement, and partnering with civil society organisations and authoritative third parties such as the IWF and Revenge Porn Helpline, was reported to be a necessary way to manage UGC and foster positive user behaviour beyond undertaking technical and human moderation. Technical moderation was seen to be only one aspect of the methods necessary to tackle harmful content across platforms and not a complete solution.

Preparing tools for parents and teachers

Online platform 4 stated in its interview that its predominantly younger user base means that they have a significant responsibility to create an online environment which not only allows for safe user interactions: the platform endeavours to educate their users on a range of harms and have developed online tools and information sources that can be accessed by parents, teachers and young users themselves. In this respect, the platform explained that 'it takes a whole community to raise a child', and that a community of parents, teachers and industry actors are all responsible to ensure that a child is looked after. This approach was echoed by online platform 8 which also provided advice online to parents, and online platform 1 (aimed at children), which provides extensive online information to parents on how they can be involved in their child's use of the platform, and of the efforts of the platform to ensure child safety.

Responsibility to users; accepting country norms

Online platform 3 noted during the interview that while it had a responsibility to its users to allow them to contribute to a community that they 'want to be part of', there was an additional responsibility to ensure that beyond the community, the platform observes the norms of the countries and regions within which they operate.

While recognising the responsibility to protect its users from harm, online platform 11 identified during its interview that there was an additional need to ensure that they do not unduly interfere with user interactions, or police speech, in a way which is overly prescriptive of their values.

Further, the platform explored the tension of moderating content when the real-world affect is a less known phenomenon. An example given is a user posting about their experiences of self-harm. The impact of allowing the visibility of such content could be manifold: the process may be of therapeutic importance to the user, though at a certain threshold and volume, could potentially lead to a normalisation of such behaviour with obviously dangerous ramifications. Consequently, when the real-world implications of permitting certain harmful content aren't wholly known, the platform can be faced with competing obligations to its users.

Online platform 7 highlighted that to manage the issue of competing obligations to its users, one approach it takes is to interrupt user behaviour (as an additional measure to taking down harmful content via technology). They reported that this helps them balance their 'responsibility to ensure expression' on the platform while managing challenges around harm.

Online platform 3 noted during the interview that while it had a responsibility to its users to allow them to contribute to a community that they 'want to be part of', there was an additional responsibility to ensure that beyond the community, the platform observes the norms of the countries and regions within which they operate.

Overview of moderation strategies used across platforms

Platform representatives on accessing the survey were asked to select the various forms of moderation strategies their platforms use to control UGC. For the purpose of this question moderation strategies were defined in the following ways:

- Pre-moderation: *The moderation of content before it has become visible to other platform users.*
- Post-moderation: *The moderation of content that is visible to other platform users immediately after submission.*
- Reactive report moderation: *The moderation of content as a response to reports made by users*

Users also had the option to select 'Other'.

Survey data revealed that all responding platforms employ a combination of pre-moderation, post moderation and reactive report moderation strategies to moderate harmful content.

In the case of both **pre**- and **post-moderation**, four platforms appear to combine internal automated tools with human moderation to assess each case. Online platform 1 reported that all visual material uploaded on the platform (during games) are pre-moderated by trained staff, before being deleted if necessary.

The advantages and limitations of **post moderation** relate to the specific purpose of the online platform. Online platform 1 instead uses a content management system to alert human moderators for inappropriate language and content for further assessment. Online platform 10 reported that live streaming leaves no room for pre-moderation, regardless of the number of moderators. This is due to the specific nature of live streaming, the format of which leaves little room for intervention through pre-moderation.

All responding platforms reported that **reactive report moderation** is commonly employed through a built-in platform system that users have access to while using the platform. A representative of a video platform particularly praised this type of moderation strategy because it provides the platform with a sense of what users perceive to be important issues. Once reported, human moderators review the content to contextualise slang, sarcasm and acronyms. An online platform stated that their user reporting menu ensures that each report is channelled to the relevant customer care queue, which is staffed by experts in that area and is actioned appropriately. It then uses technical systems to manage report queues and allocate tasks to staff members and run reports.

User reporting

All platforms give users the chance to report. All platforms present a certain standard format in terms of reporting channels and technological features ("click and report" tools). Seven out of 10 platforms allow users to report both content and users. There are four platforms that let users report more than one piece of content in one batch using the "in-product" or in-app reporting tool: this can range from pictures, comments, chats, etc.

In terms of the type of content, some platforms provide pre-defined categories of harm, while others allow users to describe and explain the nature of the reported content. For example, online platform 2 explained that to report a user:

> "*You can click on the icon and choose among the available categories of harms".*

In contrast, online platform 3 indicated:

> "*Users can report harms directly from posts or other users' profiles by choosing the preferred option*".

Eight platforms have a set of well-defined harms that the user must select when reporting, while two online platforms offer a more interactive feature by allowing users to add narratives and provide explanation.

The extent to which online platforms provide feedback to users who report content, or offer information about the review process, is quite limited. Online platform 7 stated that it notifies users once they have received the content and taken it down.

Reporting users to local authorities, law enforcement and user blocking (survey data)

When asked in the survey whether they currently report UK-based individuals that are engaged in the harms they moderate to *local authorities*, two of the four responding platforms (platforms 1 and 10) indicated that they do across all harm they moderate. One of those two platforms (online platform 10) responded further that they have reported individuals to *law enforcement authorities* in the last year for each type of harmful content it moderates. The other platform of the two indicated that it only reported users to law enforcement in the past year for content or behaviour related to hateful content and conduct and sexual content.

Online platform 9 indicated that it reported individuals to local authorities and law enforcement authorities via NCMEC in relation to child sexual abuse material and exploitation. For terrorist content, this platform would report to authorities in the event of a credible imminent threat of harm, but it does block access to all terrorist content.

The online platform that submitted the narrative response to the survey indicated a variety of ways that it engages with law enforcement and other important actors in the case of illegal activity. In the case of CSAM, they provide reports to NCMEC, and pass intelligence packages on CSAM (compiled by the platform) to overseas law enforcement via NCMEC. Regarding terrorist activity, when there is a perceived immediate threat to life, the platform will report to law enforcement.

Online platform 4 indicated that it does not report users to local authorities for any of the harms it moderates, nor does it block access from the UK to content of this nature.

All survey responding platforms declared that their platforms have procedures in place for law enforcement authorities to report content to them which is locally allegedly illegal.

Three online platforms out of the four that responded to the survey stated that they perform geo blocking of content; one survey respondent from an online platform reported that it does so in the rare circumstances of a violation of local law.

Suspending user accounts

All platforms have a policy in place that means that once their users breach their Terms of Service or Use or fail to comply with the community policy documents, their accounts risk being suspended. All platforms reserve the right to close accounts for a determinate or indeterminate period, even prohibiting users to ever open an account with them again.

One online platform stated that it in some cases it will contact the user to allow them to explain or correct their behaviour; the other platforms did not indicate whether this was an approach they also had in place.

Survey data highlighted that for all responding platforms, users found to be culpable of activities or behaviours which constitute the harms they moderate might result in the user's account being suspended. This happens in all cases at online platforms 1 and 10.

Online platform 4 highlighted the importance of context in the case of suspension: CSAM is the only harm which will also result in an account suspension, in all cases. For all other harms, the same online platform reported to make an assessment based on past user behaviour, warnings and past suspensions. This approach was echoed by online platform 7.

3.2 Capabilities of platforms to address online harms

Key messages

Economic resource allocation to addressing online harms

- Out of those platforms that responded to the survey, only online platform 10 provided its total economic resources allocated to moderation. Nevertheless, precise figures were not provided in the subsequent questions of the survey as this platform reported to still be working on releasing detailed numbers.

Human resource allocation to addressing online harms

- Most of the platforms could not provide precise estimates on the human resource allocation to tackling harms, either because moderating harm is not only undertaken by sole moderators, or because it was impossible to separate employees time into figures which were precise enough.
- Five out of 10 platforms indicated that they use subcontractors to undertake moderating responsibilities.
- None of the platforms that responded to the survey could provide the ratio between human and automated interventions.

3.2.1 Introduction

The purpose of section 3.2 is to present the economic and human resource allocation to moderating and tackling harms by each online platform.

3.2.2 Human capability

Human resource allocation to addressing online harms

Most of the platforms could not provide precise estimates on human resources allocated to handling harms during interviews because it was reported that moderating harm involves multiple teams and because it is not possible to exactly separate employee's time into figures which were precise enough to be meaningful. An online platform indicated that they would not share headcount and cost data as they felt that analysis of such data could lead to unfair comparisons to be made between small companies (like the platform) and larger market-leaders.

The first issue was captured in an interview with online platform 4, which reports that cross-functional teams are responsible for tackling harms; these include engineers responsible for developing and maintaining dedicated technology to reduce harm, PR staff and members of the customer care team. Online platform 10 spoke of how it directly thinks about moderation when new products are developed as an example of human resources allocated to these tasks which are not commonly thought to be involved.

Five platforms interviewed indicate that their approach to moderating harms involves subcontractors. This is generally done by employing a combination of internal and external staff. Only online platform 2 revealed in its interview that all moderation is done in house. The other platforms did not comment on this point.

Online platform 1 reported that harms have a different amount of resources allocated to them. More precisely, it allocates a greater number of resources to grooming, CSAM and suicide ideation. Content is also reviewed according to the priority that the platform has allocated to different harms. AI algorithms are given higher scores to identify those harms in content, is therefore reflected in the time moderators spend on them.

Five platforms out of 10 reflected on the training they provide to moderators. Trainings are provided in house at online platform 2 and are considered key to understand specific market contexts at online platform 11. Another online platform recruits former employees of the National Center for Missing Exploited Children as part of their moderation teams.

Two platforms out of ten touched upon the location, and the international dimension, of their moderating teams during interviews. Online platform 11 reported that the geographical location of their moderators is irrelevant in the case of reports which relate to image-based content; text-based content instead requires language-specific skills from moderators, which might affect their location.

Online platform 9 referred to a dedicated moderation team which looks at all harms. This platform suggests that there is likely to be a relationship between the size of the platform and whether they have harm-specific moderating teams.

Similarly, online platform 7 referred to trends in the appearance of harms and consequently the need for flexibility to efficiently tackle them, as a reason for why their moderation team works transversally across harms. On the allocation of harms to moderators to work on specific harms, online platform 11 reported that it has a dedicated team working on terrorism and some market-specific teams. These teams consider how different harms manifest themselves in different ways in different countries; consequently, they tackle the uniqueness of some harms in a given market. Being part of these teams involves knowing the language and the country context accordingly.

Regarding psychological support provided to human moderators, all five platforms that completed the survey or submitted a narrative response stated that they do provide support when humans are deployed regardless of moderation strategy or the type of harm they work on[5]. Two platforms mentioned that they have dedicated breaks for moderators and a wellbeing programme in place. Online platform 7 reported that it provides resilience training to its moderators and rotates them to avoid their exposure to disturbing content for long periods. Online platform 1 mentioned that regular feedback meetings with all support managers take place; they also have access to supervision and counselling and benefit from assistance of an independent chartered psychologist for further consultation in relation to work issues.

3.3 The incentives for platforms to tackle harms

Key messages

Current incentivisation of the platforms to tackle harm

- Reputation is a key driver when it comes to how online platforms are incentivised to tackle online harms; in particular, reputational threat interacts with the role played by external stakeholders, such as advertisers, investors and paying users (depending on the revenue model and user base of the platform).
- Non-legal regulation (including self-regulatory obligations) were identified as being more impactful in shaping a platform's approach to tackling harms than standard legal obligations.
- Competitors act as drivers within a collaborative capacity; overwhelmingly, platforms indicated that opportunities for collaboration with other platforms, through sharing of knowledge and best practice, meant that certain harms (with significant consensus around legal status and meaning) could be better targeted.
- International and national collaborative networks, forums or initiatives focused on tackling online harm such as the Global Counterterrorism Forum, WePROTECT Global Alliance or IWF were cited as valuable instances to empower platforms to make online environments safer.
- The platform's unique value will often incentivise and guide its approach to tackle harm.

Increasing incentivisation of platforms to tackle harm

- Reflecting that collaboration with competitors was a driver in this area, it was noted by platforms that more opportunities for cooperation with industry and civil society enhances a platform's capacity to tackle harm. This might include more robust frameworks (e.g., as terrorism where definitions of harm and how to tackle it are more clearly defined), more frequent conferences and more initiatives that the platform could be part of.

3.3.1 Introduction

Section 3.2 reflects on how platforms might be currently incentivised to tackle online harms, drawing from the relevant literature. It then presents how platforms indicated that they were currently incentivised to do so, the role of a platform's revenue streams in its strategies to

[5] The question of whether the platform provided psychological support to moderators wasn't covered in interviews; consequently, information presented here refers to survey responses. The implication is not that only five do so.

tackle harm, and how the platform reported that they could be further incentivised to tackle harm.

3.3.2 Current incentivisation to tackle online harms

From interviews undertaken with platform representatives, **reputation** emerged as the most common driver to creating a safe browsing environment for all platforms. An online platform stressed that a good reputation will help the platform grow and increase its profits, but that this can only be done by **building trust from a range of external stakeholders** that include advertisers, investors and users.

If the platform has an advertising-based revenue model, **advertisers** will not invest unless they consider it a brand-safe platform. Advertisers do not want to see their brands next to harmful content. Online platform 3 commented:

> *"Some of our platform's revenues are based on advertising and we saw cases in 2017 where advertisers were unhappy with their adverts shown against certain content".*

Five platforms highlighted that the role of **investors** is analogous to that of advertisers; safe platforms attract users and, as result, ensure company growth and revenues by attracting investors looking for financially healthy platforms where users have a safe online experience. Therefore, the link between **safety**, **user demand and investors** is important. Online platform 2 reported:

> *"We don't have advertisers on our app, but there is an undeniable ecosystem at work: if you don't have advertisers; you don't have revenue and without revenues you can't finance the safeguarding of your users"*

Reputation, and the role of advertisers and investors are examples of a range of external stakeholders which act as drivers for a platform to tackle online harms; as mentioned, they can be interlinked, or can operate more discretely. Regarding broader **external factors**, platforms were additionally asked whether legal and non-legal obligations, competitors and technological change incentivised them to tackle online harm.

All platforms agreed that **legal obligations** are important factors in that they must comply with the law. However, legal obligations were seen to be underpinning external factors, rather than an overly prescriptive driver that is defining in how a platform chooses to tackle harms. Given that all platforms moderate harms which are both illegal and lawful, platforms will always go beyond legal obligations when it comes to moderating harms. Online platform 11 stated that legal obligations that specifically oblige online platforms to tackle a certain type of harm[6] "*very rarely have an impact at scale*" in comparison to the volume of content tackled for being in violation of the platform's community policies. Further, three platforms specifically mentioned the ethical dilemma that arises when legal obligations emerge in certain jurisdictions which heavily impinge on freedoms of expression and association.

A commonly held view among 9 platforms is that **non-legal (including self-regulatory) obligations and measures** have been more impactful and helpful in shaping approaches to tackle online harms than legal obligations have. Examples of such impactful initiatives which have created certain non-legal obligations were mentioned across nine platforms and include: the EU Code of conduct on countering illegal hate speech online, the Technology Coalition, the Global Internet Forum to Counter Terrorism, the EU Internet Forum and the ICT coalition. Online platform 7 noted that the transnational nature of many of these industry collaborations

[6] Such as the German Network Enforcement Act (NetzDG)

is important given that many problems relating to tackling online harms are experienced across jurisdictions and traverse national government policies.

Three platforms stressed that **self-regulation** triggers dialogue and mutual monitoring between stakeholders (including platforms, governments and civil society) something which is perceived to be necessary in such a fast-changing environment. Eight out of 10 platforms stressed that dialogue and tailored solutions to how platforms tackle harms are preferable to how legal regulation can impose a ‘one-size-fits-all’ approach.

The work of **competitors** was cited during the interviews by six platforms as being an important driver to tackle harms, though in more of a collaborative than competitive way. Two online platforms highlighted that what big companies do is important for small ones, who can replicate their approaches to the extent that is possible and desirable. Online platform 2 (and a small enterprise) commented:

> *“[…] smaller companies first look to the bigger ones and see what they are doing, what they define as harmful, what their policies are. The ‘giants’ play a role in setting standards. We looked at the ToS of the top 5 to understand the differences between them”.*

Online platform 11 reported that competitors were significant for providing opportunities to iterate and learn from each other. Initiatives such as the Global Counterterrorism Forum, WePROTECT Global Alliance and IWF were cited as meaningful and effective means through which industry actors (and often competitors) come together to advance common approaches to tackle harms such as CSAM online, online platform 3 (that stated that ‘you are only as strong as your weakest link’), online platform 1 (that reported that in instances of threat to child protection on the platform, they would often inform other platforms with a similar user base), and online platform 8 (that reported on the benefits of attending industry-led forums with competitors).

A representative of online platform 11 noted that collaboration with competitors was more established when the harm was defined the same way internationally, usually by international legal standards. Where there was less consensus over what constituted a particular harm, there would be less collaboration.

In addition, some platforms pointed to their **unique value** (their unique principles, values and product) as being important drivers to tackling online harm. A platform’s unique value will guide its approach to tackling harm, whether that be because the platform is aimed at children or because the platform particularly cited the promotion of freedom of expression. Online platform 7 stated that ensuring user safety was part of the ‘DNA of our company’ and drove their ‘incentives and investment’ in tackling online harms. This tied in with their self-identified ‘unique purpose’, which seeks to strongly engage with its community of users; this platform cited that it was user feedback that highlighted the need to provide greater clarity on hate speech policies. Online platform 1 mentioned that its being aimed uniquely at children meant that child safeguarding had to be its most important responsibility.

Platform revenue streams and tackling harm

Seven[7] of the platforms derive some of their revenue from advertising and/ or banners. They all unanimously report that tackling online harms was key to protecting ad-based revenue streams. As mentioned above, tackling online harms relates directly to reputation, and without solid public reputation, advertisers would not be attracted to the platform.

[7] One platform declined to comment on revenue streams in its interview.

Online platform 10 indicated during the interview that its main sources of revenue were advertisements, subscriptions, and instore value currency. They report that tackling harms was unequivocally favourable to all three revenue streams. Online platform 1, operates as a ‘freemium’, meaning that their revenue comes from users paying for subscription services. As it is parents that pay for their children to access the platform, the need to be perceived favourably by parents, NGOs and the media is key. Without the trust of those stakeholders that online harms were being sufficiently tackled, the platform would have no revenue.

Whatever the revenue stream - ad-based, subscription-based or via in-game currency - platforms report unanimously that addressing online harms was only beneficial to preserve them.

3.3.3 Increasing incentivisation

Six out of 10 platforms highlighted that more external support would enhance moderation activities and improve effectiveness. In particular, online platform 1 stated that better collaboration and relationships with law enforcement authorities would increase their capacity to tackle illegal harms and recurrent offenders who, even though have their accounts closed, go on to open others. The need for better communication with law enforcement agencies was mentioned in the context of transnational cooperation. For example, that platform stated that in an event of grooming of a UK-based user by a perpetrator that lives outside the UK, there is a need for strong and effective cooperation between national law enforcement agencies and Interpol, something which often cannot be influenced by the platform.

Knowledge sharing and conferences to enhance cooperation in terms of technological tools or best practice examples within the industry was mentioned by five out of 10 platforms as positively improving capacity to address online harms, as opposed to each company acting separately. Likewise, more cooperation with peer companies in technological developments of algorithms and machine learning tools was mentioned by three platforms as another capacity-enhancing measure.

Online platform 4 stated that although much has been discussed within the child protection industry about ‘digital resilience’, there is insufficient external support to online platforms addressing online harms to enhance users’ resilience, so they had to bring in experts themselves. They mentioned that more discussions about strategies to tackle online harm within the industry would help.

4 Understanding effectiveness and efficiency in how platforms tackle online harms

Key messages

Perception of success

- Platforms gauge their successes in addressing online harm via a combination of metrics, analysis of reporting trends, and public perception.

Metrics to measure effectiveness and efficiency

- Three out of four platforms (that responded to the survey via the survey platform) have metrics in place to measure **effectiveness**. In all cases these were not found to vary by moderated harms.
- Only one of the four platforms (that responded to the survey via the survey platform) has metrics in place to measure **efficiency**.

Incentives, economic and technical capacity and effectiveness

- Reputational threat is the incentive that was cited as most impacting how effectively a platform addresses harm.
- While there is a link between economic and technical capacity and how effectively a platform can tackle harms, these are not the sole conditions to generate effectiveness.

4.1.1 Introduction

The purpose of section 4 is to explore how platforms perceive themselves to be successful in reducing online harms, whether they have metrics in place to measure effectiveness and efficiency, and whether there are links between incentives, economic and technical capacity and effectiveness.

4.1.2 Perceived effectiveness and efficiency

Six out of the 10 platform representatives interviewed report that they rely on some form of metrics to gauge the success of their platform in handling online harms. While the metric definitions can vary across platforms, many platforms share some common indicators. As an illustrative example, time taken to tackle harm was reported in half of the interviews conducted as a key metric to measure success.

Other indicators cited by platforms during interviews include: the number of views before harmful content is taken down from the platform; the volume of harmful content removed; and efficiencies of moderators and technology in spotting harms and minimising false negatives or positives.

According to a respondent from online platform 10, the definition and purpose of their metrics are under constant revision. Online platform 9 reported that they measure volume of infringing content as well as volume of reported content that is not infringing (false positives).

A respondent from online platform 4 highlighted how they look at trends in the volume of reports of harmful content and link success to a decrease over time. Similarly, other respondents across four platforms mentioned that they consider their efforts to enhance their public perception (regarding user safety), their engagement with their users, and their efforts towards transparency when assessing their successes to reduce online harm.

While metrics appear to be the most common way of measuring success among platforms, a representative of online platform 7 highlighted their belief that referencing statistics or benchmarks when addressing harms is not the soundest approach. The interviewee reported that a platform's success should be measured by it having in place clear policies, technologies to flag content, the best possible provision of support to users and the ability to be responsive to law enforcement when necessary.

According to the same platform, greater calls to use metrics and statistics should be taken cautiously as they could incentivise platforms to underreport on some harms. Further, in the case of a take-down request that could be perceived to affect free-speech and for which the potential harm is not immediately obvious, the platform stressed the need to carefully evaluate the request, meaning that speed of take-down would not be a wholly appropriate metric.

4.2 Metrics used by platforms to measure effectiveness

Three out of four platforms that responded to the survey have metrics in place to measure the effectiveness of moderation to tackle online harm. In all cases these are not found to vary by harms moderated. Two out of the three platforms which declared having metrics in place also provided some definitions which include a focus on 'totals', such as: total volumes, processed volumes, removed volumes. One online platform does not provide any metric definitions.

Two of the three platforms that have metrics in place to measure effectiveness declared that they have metrics in place to measure false positives. Neither of the two platforms reported being able to break down these figures by false positives in automated processes, although one of them (an online platform) can break them down in false positives in human processes.

4.3 Metrics used by platforms to measure efficiency

Only one of the four platforms which completed the survey has metrics in place to measure efficiency. These metrics look at the validity and process duration and are in place across all harms which the platform moderates. This platform also allows false positives to be measured.

Three out of the four platforms that responded to the survey report having metrics in place that track the speed of content removal (speed metrics). Online platform 1, which does not have any efficiency metrics in place, reflected on the reasons in a subsequent survey question, stating that they are working to establish them.

Looking more closely at the three platforms) with speed metrics in place, two of them are also able to provide a high-level description. Online platform 9 looks at turnover time from identification to takedown, while online platform 10 measures response time.

4.4 The relationship between incentives, capabilities and effectiveness

Incentives and the effectiveness of platforms to reduce harm

Five platform respondents cited reputation as a key driver in their approach to tackle online harms. This was highlighted during an interview with a representative of an online platform 11 who stated that reputational incentives clearly drive their spending on security measures for handling harm.

The same respondent also stressed the importance of carefully balancing these incentives with a constant iteration and feedback system to make sure that real harms are tackled and that efforts are not driven only by public perception.

Reputation and the perspective of users was reported in an interview with online platform 7. The respondent underlined how incentives are independent of government pressures; instead they are driven by user engagement, something they encourage to avoid users 'voting with their feet'. The same argument was referenced by a representative of online platform 10 who then linked user engagement with the platform with advertisement spending and the potential detrimental impacts on profits when that user relationship is threatened.

From a different perspective, online platform 4 cited reputation as one key incentive to effectively reduce harm on its site, particularly to limit its susceptibility to attacks from the media regarding its (lack of) safety.

Interestingly an interview with a representative from online platform 8 highlighted something unique which was not shared by other platforms. This platform was the only one that hinted at how the presence of harm can create perverse incentives for platforms to make money; harmful content might increase advertising and increase user consumption, generating a positive link between harm and economic incentives. These would in turn generate resources to tackle the harm itself and thus create an incentive to find the optimum level of harm on the platform.

Economic and technical capability and the effectiveness of platforms to reduce harm

Four of the platform representatives interviewed explicitly reported a link between economic and technical capability and their effectiveness to reduce online harms.

According to a respondent from online platform 1, financial resources are required to implement an effective safety strategy. This can be a challenge for start-ups in the gaming industry which might still be in the process of accumulating the relevant expertise, knowledge and financials needed to implement and improve effective procedures for handling online harms, and to establish the sound technical infrastructure to do so. Similarly, small companies might lack in-house technical capabilities and find it easier to outsource some of these activities to experts.

A similar view was reflected by a representative of online platform 7 who stressed the need to effectively deploy trust and safety resources. Given that the size of players in the industry can vary from 'giants' to start-ups, this platform remarked how it is of crucial importance not to impose on smaller companies the same standards which can be expected from larger companies which are likely to have many more resources. The same platform underlined how some skills and expertise are difficult to recruit among smaller companies and praised the work of external organisations and forums which can bring experts together. They view it as critical that government regulation should not fragment efforts and compete for the resources they have already dedicated to fighting online harm.

Regarding technical limitations, a representative of online platform 10 reflected the literature in this area regarding the moderation of live content. According to the respondent, live content cannot be pre-moderated, because of its ephemeral nature. The platform is notified that a live stream contains harmful content mid-stream; consequently, it can block the stream or if it is subsequently saved as a video format and subsequently shared, it can be tracked and removed.

A link between economic and technological capability was reported during an interview with online platform 11, who perceived that there is a public conception that the 'bigger the platform' the bigger the capability to invest hence the bigger the capacity to handle harm. These seem to relate to the fact that the size of the company as the bigger the platform the larger expectations form the public due to different capabilities to invest. The same respondent also mentioned that small companies can tackle harms without the same economic capacity of larger giants. According to the respondent this can be done by tackling harms through educating users and an overall improvement in the users' understanding of risk. For harms such as CSAM and terror some of the technology can and is shared across different platforms.

According to the same platform respondent, there are limited incentives for platforms to compete to tackle harms more effectively than others. The respondent reflected on the internet as a public good and thus as a common space to protect. This was considered as relevant; if the public begins to perceive the internet as harmful, this will affect all players in the sector regardless of the actual presence of harms on specific platforms. The respondent commented that their publishing information on their moderation policies is an example of their collaborative approach to handling harms, and that smaller companies or start-ups can take inspiration from their methods and approaches.

5 Impact of regulation and other factors on reducing online harms

5.1 Regulatory impact on reducing online harms

Key messages

Impact of regulation

- Recently adopted legislation in Australia, Germany and France points to the fact that states saw a need to set up **specific rules on specific harms** (hate speech, revenge pornography, cyberbullying and fake news) setting out important penalties for platforms to increase online safety, and step up the fight against such harmful content. **Platforms see national or international regulation as an incentive if very concrete and clear in terms of defining a specific harm**.
- Legal obligations on platforms to provide transparency reports shed light on the number of national complaints, platforms' actions and procedures to take down content. Transparency reports are however not an indicator to provide information on general online safety or information on actual harmful content present on a platform.
- Platforms perceive **self – and co-regulatory approaches** as having a higher impact on **incentivising their actions to tackle online harm**. For purposes of law enforcement such approaches have been evaluated as effective to identify harmful content and install a dialogue among involved actors (government, law enforcement, platforms, civil society) especially if continuously evaluated.

Perceived economic impacts of addressing online harms

- Platforms perceive the main costs in addressing online harms to be in developing technology, purchasing relevant tools to address harms, and hiring and training moderatos to review content.

Perceived impacts of technology and transparency reporting

- Technology is seen as positively impactful to how online platforms tackle online harm. However, it is identified by all platforms that in nearly all cases technology must be enhanced by human moderation, to contextualise and analyse content.
- For those platforms that produce transparency reports, they report that it fosters trust and openness with their users. It is not seen to incur significant costs.

5.1.1 Introduction

Section 5.1 presents an overview of the extent to which regulation has contributed to identify and tackle online harms in order to create a safer online environment for its users, as is discussed in the literature. It then provides an international regulatory review (of Germany, Australia and France) to determine whether the regulatory approaches employed in those countries have affected how online harms are tackled by online platforms. It then presents how impactful platforms understand regulation to have been in relation to their tackling online harms.

5.1.1 International regulatory review

This section explains the regulatory approaches that have been employed in Australia, Germany and France to tackle various forms of online harm and any findings which indicate that these approaches have had an impact on how online platforms have been able to address online harms. Relevant literature was further explored to highlight successes and limitations in these different approaches and which could potentially guide the UK options regarding regulation.

Germany

In 2017 the Act to improve enforcement of the law in social networks (Network Enforcement Act -hereinafter NetzDG) was adopted in Germany. The law was adopted with a view to improve social media platforms action and reaction regarding user-flagged hate crime content. The law is applicable to "*telemedia service providers which, for profit-making purposes, operate Internet platforms which are designed to enable users to exchange and share any content with other users or to make such content available to the public*". It excludes professional networks, special-interest communities, online gaming platforms and shopping websites, as well as journalistic/editorial websites. In addition, the law also excludes those Internet platforms which have fewer than two million registered users in the Federal Republic of Germany. The law requires from platforms that they remove or block access to content that is manifestly unlawful (content that can be recognised as such without additional examination) within 24 hours of receiving the complaint. For other forms of reported content platforms must decide "immediately", i.e. usually within seven days of receiving the complaint whether to delete or block content. Under certain circumstances the deadline of seven days can be exceeded. Platforms have also the option to set up a self-regulation authority which can help them to decide whether content is lawful or not.

The law does not require platforms to proactively search for unlawful content and hence is applied only to reported content. In this view, the law prescribes that platforms provide for user-friendly reporting channels. Platforms that receive more than 100 reports per year are obliged to publish a transparency report. The report shall inform about:

- Mechanisms in place to submit reports about criminal content;
- Criteria applied in deciding whether to take down/block content;
- Number of complaints filed within the reporting period, broken down according to who reported the content (hotline or user?), the reason for the complaint;
- Number of complaints within the reporting period that resulted in take down/access blocking (also broken down according to who reported the content and the reason for the complaint).

The report must appear twice a year and be published in the Federal Gazette and on the platform's website.

The following intentional or negligent failures to comply with the law will be fined:

- not reporting constitutes a regulatory offence;
- violations of the obligation to maintain an effective complaints management system (systemic approach not a specific individual complaint);
- not naming a person authorised to accept service and to receive information requests from German law enforcement authorities.

The NetzDG Act has generated prior its adoption quite a large amount of press reports concerning false positives and received criticism from various sides; while for some, the provisions did not go far enough, for others the law was contested as a tool to restrict freedom of speech online. An early evaluation carried out by the think-tank CEPS (Echikson, W & Knodt, O., 2018) looked at the six-month period after the law came into force. To date, no fine has been imposed and the expected flood of notices for takedown requests cannot be found in the data published in the transparency reports of platforms. According to interviews conducted for this study[8], the 24-hour delay of processing notices seemed manageable for the bigger companies but required important resources for the smaller platform interviewed. Against Facebook several recent lawsuits were opened after NetzDG came into force with regard to content removals. Facebook has seen contradictory rulings regarding content removal. Facebook was criticised for over-deletion of content that is legal in Germany (Reporter ohne Grenzen, 2018). Facebook's NetzDG report also informed it takes down content that is against their Community Standards (Terms of Service) first and then may check if the content also violates German national law (Facebook 2018). The legal complexity of specific cases shows that no precedent standard can be established. Each situation needs to be carefully evaluated, which requires significant human resources on the side of the platforms (even when the same content is re-uploaded). Hence, platforms seem to simplify by checking content first against their own Terms of Service.

The research has also shown that all three major platforms (Facebook, Twitter and Google) developed their own specific reporting standards that seemed to impact on the user-friendliness and ultimately on the number of notices made (Echikson, W & Knodt, O., 2018). The authors of the paper thus recommended quality standards for notice mechanisms. The same is true for the counter-notice mechanism. Here a clearing house format was recommended for disputed content leading to more transparent decision-making system (avoiding over-blocking) and a more user-friendly format to offer users a way to dispute takedown decisions (Echikson, W & Knodt, O., 2018).

The paper also found that transparency reporting standards varied among platforms, limiting possibilities to compare data. It was therefore recommended that industry standards should be set. The first reporting phase however already gave some further information regarding human resources deployed to handle the NetzDG complaints, information that was not accessible prior to the reporting obligation[9].

Possible negative side-effects observed by the authors of the evaluation was that smaller platforms are in the first years of implementation under pressure to avoid potential fines. The smaller platform clearly prioritised German notices compared to notices from other areas of the world. Another issue was that due to the new procedures and compliance requirements on hate speech introduced under NetzDG larger platforms did potentially attract less hate speech, but certain types of hate speech moved on to smaller platforms (e.g. anti-Semitic speech) (Echikson, W & Knodt, O., 2018). Finally, the NetzDG did not include services such as WhatsApp and other mobile device tools which arguably are part of the problem for sharing hate speech and should have been included (Echikson, W & Knodt, O., 2018). In January 2019, a question from the Parliament to the Government on NetzDG and the corresponding answer (Deutscher Brundestag, 2019) reveals that the Federal Ministry of Justice does not have a complete list of platforms to whom the law does apply, and that the Ministry is still evaluating which platforms do currently fall under the law and so should publish a transparency report. The Government also informed that the Federal Office for Justice (Bundesamt für Justiz - BfJ) still investigates about 800 notifications from users about potential issues of non-compliance with the NetzDG Act and no official results of these investigations have been so far concluded or provided that a fine needed to be charged to a specific platform.

[8] Interviews were conducted with Facebook, Google, Twitter and Change.org

[9] In the case of Facebook, the report revealed that it has 64 staff members working on NetzDG notices, which amounted in the first six months (January – June 2018) to 1,704 pieces of content. For Google, 100 staff members work solely on NetzDG notices which amounted in the same period to 241,827. For Twitter it is 50 staff members for NetzDG notices, of which Twitter received 260,000 notices in the same period. For Change.org - a small platform - 4 staff members were dealing with notices – 520 related to hate speech in the same period (but only working weekdays, not weekends. Then global staff took over).

Australia

In 2015, Australia enacted the Enhancing the Online Safety for Children Act. The Act at the time established an e-Commissioner for child safety online to reduce socially undesirable behaviour of cyber-bullying of children. The law introduced a notification mechanism for cyberbullying content material targeting an Australian child to be taken down from all large social media websites; it also introduced penalties for non-compliance. The Act provides the e-Commissioner with the power to investigate complaints about serious cyberbullying material targeted at an Australian child.

The notification mechanism is based on a two-tiered scheme: in Tier 1 are those social media services that participate on a voluntary/co-operative basis with the Commissioner; Tier 2 includes large social media services that are declared by the Minister for Communication that are subject to legally binding notices from the Commissioner (content has to be taken down as a consequence) and civil penalties (fines) in case of non-compliance with the notice. The Act considers it to be good practice that social media services have a complaints management system, terms of use which sufficiently prohibit cyber-bullying material and a contact point for the Commissioner to refer complaints that users consider have not been adequately dealt with for the removal of cyberbullying material from participating social media services. The mechanism means that the victim first notifies the user posting harmful content. Only in cases of non-reaction or non-compliance can the victim defer the case to the e-Commissioner, which then has 48 hours to intervene with industry to get the content taken down. The Commission can also issue a notice to individuals who post cyberbullying material and request the take down of that material. The notice will also include a requirement that the end-user posting that content must apologise in the format as stated in that notice and within a specific delay. The Commissioner monitors the process and in case of contravention by the user the Commissioner can issue a formal warning and/or an injunction to do so.

In 2018, the Enhancing Online Safety (Non-consensual Sharing of Intimate Images) Bill amended the 2015 Act. The amendments slightly modified the title of the Act to Enhancing Online Safety Act 2015, broadening the powers of the e-Commissioner and adding the administration of a notification mechanism on non-consensual sharing of intimate images or videos (also known as “revenge porn”), which is not victim restricted to children. The Commissioner can send a notice for removal for such imagery to social media services, the end-user of the social media service, in addition to the hosting service provider. In case of non-compliance of the end user, the Commissioner can impose civil penalties (this can be prison sentencing or fines).

In addition to the Enhancing Online Safety Act, other rules are in place that categorise illegal or potentially illegal content. There are various schemes under the Broadcasting Services Act from 1992 that are implemented via self-regulatory industry codes of practices. The Commissioner has additional investigatory and notification powers stemming from these rules. In the case of content that is of a serious nature, the Commissioner can send a notice to the legal enforcement authority and send a notice to the internet service provider according to standards of the industry code of practice so that the provider can deal with the content accordingly.

It shall also be mentioned that in 2018 a Bill (Assistance and Access Bill) was passed that provided for additional cooperation with IT companies helping law enforcement to get access to data of convicted material – this can involve getting access to passwords, accessing encrypted material or asking companies to develop encryption for police operations.

The Australian government started in 2018 an official review to assess impact of the current regulatory framework on tackling illegal or allegedly illegal content online. The report published in February 2019 (Briggs, 2019) highlights that technological advances considerably influenced changes as to how illegal material is exchanged (specifically developments concerning mobile devices not sufficiently considered by regulation). Criminals can respond rapidly and change methods of sharing defying conventional notice and compliance mechanisms. The current system remains relatively uncoordinated and the regulatory framework is fragmented in Australia (applicability of rules to online devices, types of platforms). The report criticises that the system is based on a reactive model when damage has already been done to vulnerable internet users and should change to a model that regulates safety measures upfront and by design (e.g. legislation should require proactive identification of content). Benchmarks for internet safety should be set much higher as is currently the case, requiring only common minimum safety standards. The report also highlights that enforcement can only be effective if implemented by companies, law enforcement and the online community.

On the other hand, the report also highlights that the official (legal) complaint mechanism as administered by the Commissioner has proven effective and was fully complied with (Brigg, 2019) . This meant that regarding CSAM for example, no content was taken down between 2016-2018 by the Commissioner (eSafety Office) and content was increasingly hosted outside Australia (Briggs, 2019). This points again to the situation already highlighted in the previous section (5.1.1.) regarding the UK approach on CSAM. Thanks to effective national regulatory strategies, harmful content can be effectively tackled in one's own jurisdiction. The Commissioner's engagement with international law enforcement (e.g. INTERPOL) and hotline networks such as INHOPE contributes to take-down content hosted offshore (outside Australia). Even if not yet perfect, a more coordinated international approach seems to point to the direction that countries can overcome the transnational dimension of cyberspace. Here, technology may also contribute to more effective outcomes. The UK Home Office invested in the Arachnid technology a project operated and developed by the Canadian Centre for Child Protection (Canadian hotline). The technology crawls the web for content that has been identified by the Canadian hotline and the US hotline (National Centre for Missing and Exploited Child (NCMEC)) and confirmed as CSAM. The technology can be deployed across websites, forums, chat services and newsgroups to instantaneously detect illegal content. The technology also sends the notice for take-down to the service provider that hosts the image (if the provider is in the US or Canada, for other jurisdictions it sends a notice to the responsible hotline of that jurisdiction). Hosting service providers in some countries (like the US) have the obligation to notify on that basis also the competent law enforcement authority (however this notice is not automatically provided for by the technology) (UK Home Office, 2017).

Regarding industry codes of practice, the report evaluated the ones in place (four in total) as out of date. While good relationships between the Commissioner and the industry is maintained the legal framework is not flexible enough to leave more space for the industry on the one hand and on the other hand in cases that evidence shows that industry codes are failing to be effective, that also the Commissioner can replace the practice by providing for new independent standards that should be picked up by industry. Relying on good-will alone has not proven to work out and self-regulatory systems needs similar monitoring and evaluation as any other legislation (Briggs, 2019). The recommendation made in the report was to develop a single fit for purpose technology neutral code of practice setting general behaviour benchmarks and compliance for online safety for industry and end-users. In addition, the Commissioner should also set (an) industry standard(s) to remain flexible to respond to evolving new types of harmful content. Codes of practice were considered as necessary to provide for flexibility and practical implementation on the ground (Briggs, 2019).

Finally, the report points out that data collection to monitor effectiveness and impact of regulation is important when regulating cyberspace. Transparency data should be provided by industry as well as law enforcement or the Commissioner to further enhance understanding of what is happening in the online space (Briggs, 2019).

France

In December 2018, France adopted new legislation[10] to tackle “fake news”, in particular during election periods. The law modifies the Electoral Code (as well as other laws) imposing on platforms during the three months preceding elections (and until the elections take place) a requirement to provide to the citizens fair, clear and transparent information on:

- the identity of the natural person or on the corporate name, registered office and corporate purpose of the legal person and of the person on whose behalf, if any, it has declared that it is acting, who pays the platform remuneration in return for promoting information content related to a debate in the general interest;
- the use of their personal data in the context of the promotion of information content related to a debate of general interest;
- on the remuneration received in return for the promotion of such information content when the amount exceeds a specified threshold.

Platforms shall provide all this information above, regularly updated and aggregated, in a register made available to the public by electronic means during the pre-election period (three months prior the first month of set date of general elections and until the date of the actual ballot).

In general, platforms need to set up a notice mechanism that is easily accessible to users to report false information. Platforms shall also implement measures to prevent the dissemination of false information likely to disturb public order or alter the sincerity of one of the votes (also outside the defined election period). Such measures include:

- transparency reporting on algorithms used to promote information;
- promote content from companies and news agencies;
- delete accounts that massively propagate false information;
- inform users about the nature, origin and distributor of the content;
- implement media education measures.

Platforms need to inform the French Audiovisual Council (Conseil supérieur de l'audiovisuel - hereinafter the Council) annually, in the form of a report, the measures implemented; in addition to information on the resources devoted to their implementation and specifying the procedures for implementing these. The Council monitors the application of the rules and contributes to the fight against the dissemination of false information by sending specific recommendations to platforms to improve the tackling of false information. The Council will also publish a periodic review on the application and effectiveness of the measures taken by the platforms. To this end it has the right to collect all information necessary from the platforms. To this end, platforms need to nominate a legal representative that responds to such requests.

Those online platforms that use algorithms for recommending, classifying or referencing information content related to a debate of general interest are obliged to publish aggregate statistics on the functioning of this algorithm. The publication must provide information on:

- share of direct access, without the use of recommendation, ranking or referencing algorithms;
- indirect access shares attributable to the algorithm and the platform’s internal search engine, as well as other algorithms that were used to access the content;

These statistics need to be easily accessible to users in a free and open format.

[10] Law No 2018-1202 of December 22, 2018 relating to the fight against the manipulation of information.

The law also provides for a co-regulatory element. Platforms can conclude cooperation agreements with news agencies, publishers of press publications, online news services, organisations representing journalists and any other organisation to implement measures to fight against dissemination of false information.

In the case that "inaccurate or misleading allegations or imputations of a fact likely to alter the sincerity of the forthcoming election are deliberately, artificially or automatically disseminated in large numbers through an online public communication service", a judge may at the request of the public prosecutor (or any candidate, political party, group or any person having an interest to act) prescribe that the platform, or any other person concerned stop disseminating the misleading information[11]. The judge hearing the complaint will take measures within 48 hours (first instance and appeal).

If rules are not respected in the period prior to elections, fines of up to 75 000 EUR and/ or imprisonment of a maximum of one year can be imposed on platforms (on their legal representative). Additionally, users that posted the information can be sanctioned. The Audiovisual Council can also suspend agreements between platforms and news distributors, in particular if these are at the origin of a foreign state.

The law was heatedly debated by stakeholders. One of the arguments of the opponents of the law was that a regulatory initiative should be taken at the European level, rather than by one state. Platforms (Google in a debate at the Senate) highlighted the difficulty to differentiate between content that refers to an information website (news) or a website that contains content of informative nature (Sénat, 2019b) rendering the boundaries and the applicability of the law unclear. The law was adopted by the General Assembly but rejected in the Senate (in the French procedure if the text is still not adopted after two reconciliation procedures the text is finally debated in the Generally Assembly in the third stage which can lead to adoption). The Senate rejected the text because the law was prepared without in-depth evaluation or impact assessment. In addition, the Senate considered the law as not effective to fight against the actual risks related to fake news and seen as danger for the freedom of expression online (Sénat, 2019b). The originally proposed legislative text was further amended in the adoption procedure. Specifically, the transparency obligations on algorithms and financing of information campaigns were introduced and further refined during the procedure (Rees 2019).

5.1.2 Current perspective on regulation

All 10 platforms revealed in interviews that while they see legal regulation as important to clearly identify illegal content or behaviour and to establish instances of cybercrime, that self- or co-regulatory approaches had more impact in shaping a platform's approach to tackle online harm. Self-and or co-regulatory approaches were understood by all 10 platforms to be a means to trigger a continuing dialogue among stakeholders involved (platforms, civil society organisations, governments, law enforcement). In this respect, there seems to be strong cooperation between platforms. Interview responses suggest that when competitors engage in self-regulatory initiatives which are industry-led, better, informed consensus among platforms can emerge on different harms, and how they can be tackled at scale. Approaches like the European Code of Conduct on countering illegal hate speech online, the Global Counterterrorism Forum and the EU Internet Forum (on Fighting Terrorism Online) were as seen as good practice by nine platforms as these were setting international or EU-wide common minimum definitions and approaches for identifying harmful content and acting on it. Two platforms cited the Advertising Standard Agency (ASA) as having been impactful.

Regarding UK or other national regulation (when a platform was based outside the UK), no specific regulation was highlighted as having been impactful, beyond specifying definitions of

[11] Modified Article Art. L. 163-2.-I of the Electoral Code

illegal content or behaviour. Six platforms stated that while they observe all legal obligations, they do not guide their approaches to tackle harms. A perceived limitation of national legislation is that it isn't adaptable to new harmful phenomena, as raised by online platform 4. Regarding future national regulation, it was seen by online platform 10 as risky to impose a 'one-size-fits-all' model to tackling harms which would be incompatible, or ill-fitting, with different platform types: the potential risk being that platforms move their operating bases to a different jurisdiction. Echoing this worry, online platform 3 highlighted that it would be preferable if regulation took a systems-based approach that considered the policies and systems platforms had in place to tackle harms, rather than imposing fines on an individual basis. Online platform 11 cited the German NetzDG law to demonstrate that legal obligations rarely have an impact 'at scale'. For this platform, the NetzDG, was seen to have directly impacted the take-down of some pieces of content – but when this was compared with content taken down which violated their community policies, the difference was enormous.

The General Data Protection Regulation 2016/679 (GDPR) was highlighted by four platforms as a helpful to protect users and improve online safety, while providing a framework across several countries. The E-Commerce Directive was additionally cited by two online platforms as being very influential through its establishing the guidelines and defences which allow a platform to remain open, but which ensures that companies are responsible for content as they become aware of it. Online platform 1 mentioned the Children's Online Privacy Protection Act (COPPA), which is US law, had been important in shaping their approach.

5.2 Economic impacts of addressing online harms on platforms

5.2.1 Introduction

This section presents the perceived economic impacts of addressing online harms, as revealed by platforms during interviews. They broadly include the cost of technological tools and in hiring and training human moderators.

5.2.2 Perceived economic impacts of addressing online harms

Despite not being able to provide substantial data on the financial allocation to tackling online harms, platforms did elaborate on the allocation of human resources, albeit descriptively. Many platforms report that the way in which roles and responsibilities are shared across several teams makes it impossible to allocate or provide specific figures on human resource allocation dedicated to tackling online harms. Besides hiring moderators, six platforms suggested that highly qualified engineers or tech developers were the most expensive costs in this area, given their high level of qualification. Online platform 9 explicitly quoted costs for taking care of the wellbeing of their moderators, who all undergo a wellbeing programme provided by dedicated wellbeing consultants.

Elaborating on additional costs, online platform 1 reported technology to be the most significant drain on resources. This is connected to the cost of setting up and running AI tools. The same argument was reflected by online platform 6 which discussed the costs associated with running servers and associated facilities, and by online platform 7, which also mentioned the costs for technological licenses.

The same view was shared during an interview with a representative of online platform 4, who claimed that technological development is among the key costs for proactive handling harms. According to the same platform, and to online platform 1, the same does not apply to technologies for implementing effective reporting channels, where costs for human resources are more predominant given the number of moderators employed.

Online platform 7 cited the developing of systems to record and disclose reports, and the human resources from project teams working on producing the transparency report, as adding to the costs involved in transparency reporting.

Online platform 9 reflected on the costs associated with tackling online harms and found that the knowledge exchange and technology sharing in the industry helps to keep costs down. They also praised the efforts and benefits obtained through the Technology Coalition.

5.3 Impact of technological tools to proactively detect harms

5.3.1 Introduction

This section presents a short overview of the ways in which technological tools are enabling online platforms to tackle a range of online harms. It then reflects on the role played by technological tools in assisting platforms to proactively detect harms, as reporting in interviews.

5.3.2 Perceived impact of technological tools to proactively detect harms

Positive impact of technological tools

In interviews, technological development was cited by all platforms as being beneficial to both proactive and reactive processes in place to tackle harms. Within a proactive context, the creation of AI classifiers mean that algorithms are a central part of harmful content moderation, while reactive technology can enable the platform to more effectively respond to user reporting. Online platform 11 reported that technological change alongside user demand were the most impactful external measures in enabling the platform to tackle online harms; online platform 8 reiterated that technology has been 'critical' in this area.

Technology was seen by platforms to be a highly useful means of ordering reports that are flagged by the platform's users to support human moderators. Online platform 1 indicated that around 90% of the reports it receives from its users do not actually violate its community policies; thus, it has tools in place that rank reports according to their severity, meaning moderators will first address those relative to child sexual exploitation, suicide ideation and grooming (followed by bullying).

Technology and context

Echoing the literature in this area, all platforms stated that despite the benefits that technology has brought them in moderating online harms, there is nearly always a need for some degree of human review. Harms for which context is necessary – cyberbullying and hate speech for example – require a greater need of human moderation than harms such as CSAM or terrorism content, which is largely image based and whose distribution online is illegal in almost all cases. Consequently, this can mean that the platform won't have enough content to train a classifier to tackle that harm, as highlighted by online platform 3. Online platform 4 with a young user base mentioned that technical and human review are limited in detecting what might be happening in the home of a user posting harmful content or behaving in a way that might constitute bullying. Consequently, the limitations of technology to understand the potentially damaging domestic/ familial context of a young user can have implications for the appropriateness of the action taken: terming a user a 'bully' can potentially have implications which in the case of young people must be considered.

Online platform 7 echoed that technological tools are unsuitable for those harms whose harms requires an assessment on whether behaviour is repeated or sustained (such as harassment or bullying), or which requires the offline engagement of others to gather facts related to the issue (in the case of impersonation). Considering these technological limitations, this platform stated that it places emphasis on fostering 'positive and supportive' online communities (which encourages users to report harmful behaviour and content they encountered on the platform), and to be able to provide support to vulnerable users.

Over-censorship

Online platform 10 responded that too much automated moderation could damage user experience by over-detection and the generation of false positives. The challenge of over-censorship was highlighted by online platform 3 as being both immediately problematic (removing non-violative content can be an annoyance for the platform user) and more widely so by potentially corroding the trust of its user base for undermining their freedom of expression.

5.4 Transparency reporting

5.4.1 Introduction

This section gives an overview of the role that transparency reporting can play in how online platforms tackle online harms, and where the obligation for platforms to be transparent can identified in industry-led principles. It then reflects on the role of transparency reporting for the participating platforms, as reporting in interviews.

5.4.2 Perceived impact of transparency reporting in tackling harms

Reasons for undertaking transparency reporting and perceived costs

Of those platforms that do undertake transparency reporting, the reasons cited for doing so predominantly relate to a want to foster trust, reputation and openness.

Online platform 11 elaborated further that there was a civil rights dimension behind their reasons for producing transparency reporting; as their transparency report presents decisions made for taking down UGC that violates their policies and not the law, they feel an accountability to explain publicly why they do so. This viewpoint was reflected by another online platform. Online platform 8 highlighted that producing transparency reports was 'the right thing to do'.

Online platform 4 indicated that there were no significant **costs** involved in the transparency reporting they undertake. In developing their technological tools, they 'think ahead to reporting' so that data can be easily extracted and presented. Of the other platforms that reflected on costs involved in transparency reporting, online platform 7 mentioned the mentioned the systems in place to record and disclose reports. Online platform 6 simply stated that transparency reporting was a 'necessary and expected cost' with 'a lot of staff' working on it.

For two the platforms which elaborated on their not producing transparency reports, online platform 1 indicated that as a company this was not a current priority, but that in the future they would endeavour to produce them; online platform 10 responded that they never produced transparency reports as they never perceived the need.

6 Conclusions

Key research objective: to understand what different platforms define as harms on their platforms

Platforms prioritise harms based on the interaction between external factors, the inherent format of the harm, and legality. For each harm, this interaction is unique.

Platforms prioritise and define harms based on different factors. Those factors that originate in the external environment are *public opinion and social norms* regarding the severity of a harm, and the *institutional pressure* that exists to tackle that harm. These two external factors interact with each other and are usually mutually connected; when a harm violates public opinion, or social norms, there is usually accompanying *institutional pressure* – such as from governments or law enforcement. This external pressure acts as a driver affecting which harms are prioritised and defined as such by the platform. Prioritisation is further affected by the harm's content type. For example, from a technical perspective, image-based content can be easier to identify and remove, by using hashing technology, for example. Word-based content can be more difficult as the severity of such harms will often be based on their context, making their identification and removal more difficult.

Legality does impact which harms are defined and prioritised by platforms. No platform included in this study permits illegal content or behaviour on their networks. Further, if a harm is prohibited in law, it offers a legal definition which can make the process of identifying it less complicated for the platform. For illegal harms such as **endangerment of children**, and to a certain extent **terrorism-related content**, national and internationally agreed *definitions* and industry-developed *technical approaches* combine with significant *public opinion* and *institutional pressure* for platforms to tackle these harms to ensure that these are highly prioritised as defined harms across platforms. Further, the fact that many forms of these mentioned harms are *image-based*, further interacts with these other factors to ensure that these are harms which are commonly defined and prioritised across the majority of platforms in this study.

Not all factors interact at one time to affect prioritisation. The three most commonly defined and prioritised thematic harms were adult sexual content, endangerment of children and violent content and conduct. Endangerment of children captures a large amount of illegal activity, while dangerous content and conduct includes terrorism- a harm with in its broad conceptualisation is prohibited in national and international legislation. However, adult sexual content as a harm contains a considerable amount of adult pornography. While not illegal (excluding some forms of extreme pornography), this is still a widely defined and prioritised harm across platforms because social norms and institutional pressure are still significant enough that platforms will endeavour to remove it from their networks.

Key research objective: to understand the incentives, capabilities and methods of different platforms to address online harms, including technical and economic capabilities

Reputation is the central driver to how platforms address online harms

Reputation was reported as the central driver affecting how platforms are incentivised to tackle online harms. Reputational threat interacts with important external stakeholders - advertisers, investors and users (both paying and non-paying). When a platform's reputation is damaged, platforms perceive that those external stakeholders will take their investment and custom to other platforms. When a platform has a poor reputation for online safety and tackling harms, it tends to have a negative public perception, and may face institutional pressure to enhance its processes in doing so.

Self-regulation is reported to be more incentivising that legal obligations

When it comes to the platform's approach to tackling online harm, self-regulation was reported to be more incentivising than legal regulation, which platforms understand can be overly prescriptive. While platforms welcome established and robust definitions surrounding harm (which can arise because of industry led initiatives and collaboration, or which are defined in law), it was reported that a 'one-size-fits-all' approach (which would direct the specific action taken and methods employed for all platforms) would prove inflexible to new and emerging harms. Additionally, it was reported that such regulation would stymie progress that is made through industry-led initiatives.

In terms of existing legal obligations, platforms report that their policies on harm always go beyond what content and conduct is prohibited in law.

Platforms are incentivised to cooperate with competitors

Competitors act as drivers within a collaborative capacity. Platforms reflected that the sharing of best practice and knowledge with other platforms was an incentivising factor for platforms to tackle certain harms - this was particularly true in the case of harms where there is significant consensus around the real-world impact of that harm, legal status, and definition. When it comes to tackling online harms, platforms don't tend to compete against each other; it is understood that online harms can have a negative impact for a platform even when the harm is hosted on another network.

Technology is of great importance, but human review is nearly always necessary

While technological development has enabled platforms to tackle certain harms more effectively – particularly CSAM and terrorism content - human review is usually always necessary. This is especially true when determining the level of harm posed by content or behaviour requires analysis of contextual factors, or when harmful behaviour or content is determined through its being repeated or sustained (such as bullying or harassment). As a result, platforms develop other methods to tackle the online harms which pose a threat to their users. This includes engaging with internal and external experts in the area and developing tools to educate users, teachers and parents.

Key research objective: to measure how effective these incentives and capabilities have been at reducing harm

Effectiveness is impacted by economic and technical capabilities, but these are not the only conditions necessary to effectively tackle online harm

Economic and technical capabilities are factors which enable how effectively a platform can tackle online harm. For smaller companies, resources are finite and there is a limit to how much certain platforms will be able to invest in technology and/ or staff to work towards identifying and removing harm: this is especially true for start-ups and small platforms. However, while economic and technological capability can have a constraining effect on how effectively a platform can address online harms, they do not form a total barrier to effective work in this area. Smaller companies can often leverage the technology that has been developed by larger companies, and invest in non-technological efforts to address harm, such as developing less expensive educational tools and increasing awareness-raising efforts around certain harms.

Reputation, user-bases, advertisers and investors are incentives which when strong, increase the effectiveness of a platform to tackle online harms by guiding their approaches. While effectiveness is increased when incentives are greater, economic and technical capacity will restrain how effectively a platform is able to tackle harms (though is not a total barrier).

Key research objective: to understand the potential impact of regulation on these platforms

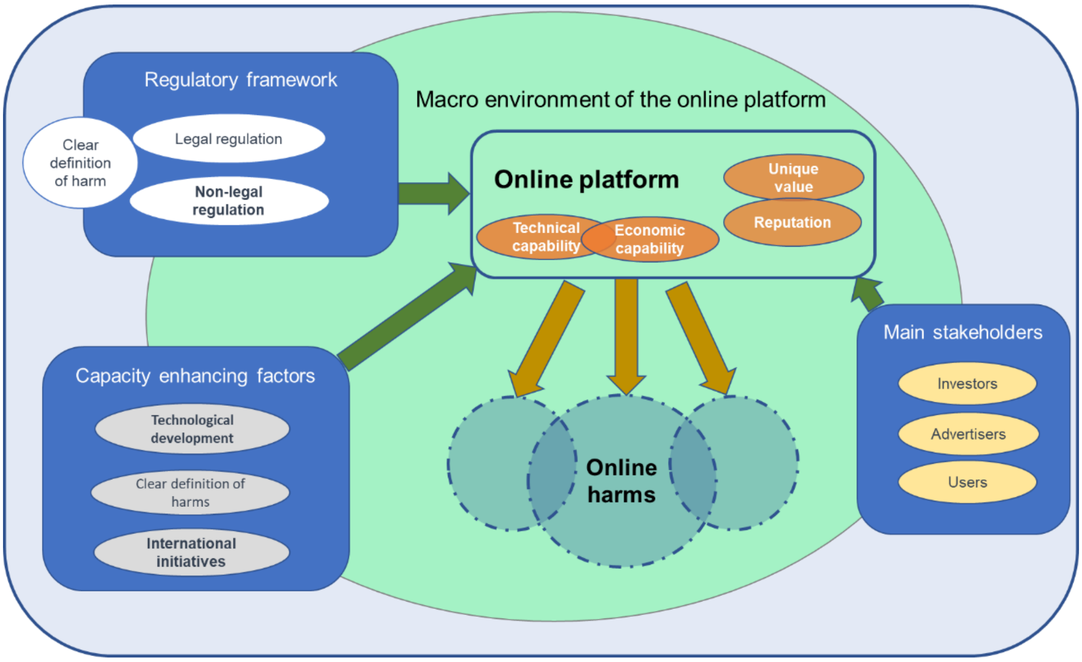
Platforms perceive that self-regulation allows them to more effectively tackle online harms than legal obligations

Platforms report that self-regulation is positively impactful to the approaches taken by platforms to tackle online harms as they encourage constant iteration and knowledge sharing in this area, whilst also helping ensure that platforms are agile enough to tackle harms which are new or emerging. Further, legal regulation that prescribes that harms are moderated in a certain way with a defined amount of resources, might cause competition between resources when a platform must redirect resources already allocated successfully to moderating a certain harm.

6.1.1 Business model

Based on research outcomes, a business model has been developed to establish a conceptual overview of the main business processes relevant to how an online platform operates to tackle online harms. It attempts to explain how those main processes are linked or dependent. Within the framework, there is a focus on the incentives, capabilities and barriers which impact how an online platform can tackle online harms. The model considers how platform operations are affected by the platforms’ environment, including the legal/non-legal regulations which framed business processes. Figure 6.1 presents the business model.

Figure 6.1 Business model



It is important to highlight the heterogeneity of the type of platforms in this market. Nonetheless the components of the business model which are outlined below interact uniquely in the case of each online platform, to affect how it operates to tackle a range of online harms.

The online platform

Platforms in this market evidence significant heterogeneity regarding their size, target users, revenue models and cost structure. They operate in niche markets; there is not significant competition between them because of the uniqueness of the different products and services they offer; as a result, they each serve the needs of different user bases. Each platform has a different revenue model. These range from ad-based, subscription, investor-led, or a combination. The size of their user base also varies according to the market.

Incentives

The main drivers to tackle online harm derive from an interplay between a platform's reputation and the influence of its stakeholders. The type of stakeholder depends on the revenue model, but for all of them, the logic is similar. For instance, if has a platform has an ad-based revenue model, advertisers will want to see their brands positioned alongside safe content; likewise, if a large proportion of revenue derives from investors, platforms need to assure them that the users are guaranteed a safe experience, as this in turn secures the platform's sustainable growth.

Capabilities (internal and external)

Internal factors affect the companies' own capabilities, while external factors can enhance those capability to tackle online harm.

The main driver behind a platforms' tackling of online harms is having a technologically-enhanced team of moderators working within the company. Both technology and human moderation are complementary: machine learning algorithms (artificial intelligence - AI) are not perfect tools, and human review is necessary to contextualise data and interpret harms which might be subtle or complex, and to minimise false positives. Tackling harms online requires substantial resources; larger companies can afford to establish an IT infrastructure capable of doing so, develop their own technological tools and hire engineers at a greater extent than smaller companies can.

As mentioned, reputation is a key driver in determining the strategies a platform has in place to tackle online harm. Aligned with reputation is a platform's unique value. Its unique value can be directly associated with reputation in a way that fosters a more aggressive approach to tackling online harms, when for example, the platform's services are aimed at children. In this case, its unique value (offering say, a particular product for children) is twinned with its reputation and in order for both to be maintained certain online harms must be tackled. In the case of a platform whose unique value and reputation lies partly in its championing freedom of speech and association, then a desire to preserve these both might result in a less aggressive approach taken by the platform to tackle harms. Thus, a platform's user value interacts with reputation quite distinctly in each case.

Within **the external environment**, international and national initiatives enhance the capacity of the platform to address online harms: this is especially true for smaller companies that for a lack of technical or financial capacity, need greater support to develop their approaches to safety. International initiatives or other activity that aims to increase cooperation between platforms (and other actors) offer opportunities for knowledge and best practice sharing and can provide clear definitions for certain harms. Clear definitions, by offering clarity on a phenomenon, enable a platform to more strategically and confidently.

Platform operations are framed by a regulatory framework, comprising both legal and non-legal regulation (including self-regulation). The majority of online platforms in scope of this study indicated almost exclusively they operated beyond what they were obliged by legally; non-legal regulations, such as self-regulated industry codes of conduct, were reported by platforms to have been more instrumental in shaping their approaches to tackling online harms.

Barriers (mostly external)

Much as clear definitions enhance operations in place to tackle harm, ambiguous (or an absence of) definitions surrounding a certain harm can negatively affect how the platform is able to address it. Further barriers include limitations on resources (financial and human); and when there is tension between tackling an online harm, and preserving the platform's user value. Definitions tend to come from regulatory frameworks.

Annex 1 Literature review

Introduction

The purpose of the literature review is to attempt to answer (slightly modified) RQs and sub-questions. The review is presented by KRO. While certain areas yielded a greater number of relevant sources, other were less fruitful. The results are presented here.

Key research objective: to understand what different platforms define as harms on their platforms.

Harm type

The range of harms that platform users can be exposed to online are numerous and constantly evolving. As one example, the internet has long been used as an environment within which to facilitate the sexual exploitation of children and distribution of child sexual abuse material (CSAM). Technological development offers new challenges to tackle child sexual exploitation (CSE) and the spread of CSAM online; on-demand live streaming of abuse is leading to the rise in the volume of CSAM online (Europol, 2016), while the use of cryptocurrencies can be used by offenders to pay for CSAM presents another future challenge.

Grooming is anticipated to increase as online platforms widen their social functionalities (instant messaging, photo video and music sharing, etc) and thus opportunities for young people to be targeted by perpetrators (Baines, V., 2008; European Commission, 2013) The online sphere further increases the risk of users to 'sextortion' – a form of blackmail where sexual content is used to extort sexual favours and/ or money from a victim (European Commission, 2013; Wolak, J, et al).

Cyberbullying is another risk which poses a particular harm to children and young people. Its recent 'upsurge' has negative implications for the wellbeing and emotional heath of those users experiencing it and can constitute a variety of forms (Cowie, H., 2013). Associations between cyberbullying involvement and self-harm and suicidal behaviour have been identified (John, A., et al, 2018), flagging two additional harms which vulnerable communities of users can be exposed to online.

Nude imagery is another widely banned 'harm' on social media platforms, the management of which differs depending on the platform. As a moderated harm, it has been the source of some contention when breastfeeding or post-mastectomy photographs have been removed, while campaigners argue for parity between rules imposed by platforms on male and female toplessness (Onlinecensorship.org, 2016).

Extremist content and related behaviour poses another increasing risk to platform users. As an indication of prevalence, it was reported that in 2014 there were at least 45, 000 pro-ISIS accounts on Twitter. Further, video hosting platforms have become a means to radicalise users and preach hate: the Counter Extremism Project identified over 80 European and US Islamic extremists that had watched and been influenced by the same al-Qaeda extremist online (Counter Extremism Project). Platforms are under increasing pressure to moderate other forms of extremist content and behaviour, such as that which perpetuates a far-right extremist and harmful ideology. Moderation of this form on content can be difficult, especially when the platform models its services around openness and free speech (Data & Society, 2018).

In defining their policies to capture some or all these online threats, online platforms must also negotiate their role as 'gatekeeper' in balancing the extent to which their users exercise freedom of expression and association (Lynskey, O., 2017). They must also decide exactly what content and behaviour constitutes these harms, which presents a particular challenge given the vast cultural, linguistic and social differences the platform must accommodate (Onlinecensorship.org, 2016).

Key research objective: to understand the incentives, capabilities and methods of different platforms to address online harms, including technical and economic capabilities.

Content moderation

Content moderation is the process by which an online platform monitors the user generated content (UGC) on its site against a set of defined policies and rules. It involves the 'listening, escalating and responding to inappropriate UGC' (Crisp, 2019) and comprises different methods and approaches which can be employed together or alone, depending on the nature of the inappropriate UGC and the stated policies and rules of the hosting platform. There are a range of content moderation types: pre-moderation, post-moderation, reactive moderation and distributed moderation (Social Media Today, 2010). Pre-moderation involves the review of UGC before it becomes visible to other users on the platform. Post-moderation consists of the review of UGC once it has been posted to the platform and is already visible to other users. Reactive moderation allows platform users to identify content they deem to be a breach of the platform's official term of use or policies, or inappropriate for another reason. Distributed moderation is a form of moderation where a user's ranking or grading of content is aggregated to determine the visibility of that content (Mills, 2013).

Incentives

The incentives for online platforms to tackle online harms can be manifold: internal or external, ranging from peer effect, self-regulatory and industry standards to more legally binding measures. Legal frameworks and approaches to regulate content varies by countries. Some countries apply strict regulatory controls on Internet and other related service providers, such as filtering or blocking access to content by technology, while others rely on a more self-regulatory approach that is further framed by the information sector (UNODOC, 2012).

Regarding self-regulation of the UK social networking sector, (Haynes, et al. 2016) highlighted that several leading players in the industry understood self-regulation to be a strong and incentivising driver to comply with industry standards, also because it affects both codes of conduct and legislation, having a multiplier effect. Further, as discussed in a 2009 Oxford Internet Institute report, one of the advantages of self-regulation was identified as being that because stakeholders are incentivised to commit to certain standards when they have embraced them. The decisive factor is that public and industry's objectives are aligned (Wales, 2009).

In terms of legal regulation, the UK has the Data Protection Act 2018, which focusses on access to personal data. This legislation was framed by the General Data Protection Regulation, which came into force in May 2018.

Additional sources of reference in the UK include the Human Rights Act 1998 (protecting individual rights), the Communications Act 2003 (focused on specific aspects of online communication), and the Consumer Protection Act 1987 (offering a framework in the advertising industry). The latter also impacts the digital advertising sector, which is of relevance to certain online platforms (OII, 2009).

Within the EU context, several initiatives have been put forward as early as 1999 to make the Internet a safer environment. The "Safer Internet Programme" was established in 1999 by the European Commission to promote the safer use of the Internet by educating users, whilst fighting against illegal content. Children were the main target population of this initiative;

currently, the programme has expanded to cover new emerging issues such as grooming and cyberbullying (UNICEF, 2011).

Other EU initiatives are examples of self-regulation. The ‘European Framework for Safer Mobile Use by Younger Teenagers and Children’ (European Commission, 2018b) was signed in February 2007 by leading mobile operators and content providers across the EU. As a result of this, codes of conduct in relation to children as victims of online harm were in place in 25 EU member States as of June 2010. Signatory members committed to principles and measures that captured “access control for adult content, awareness-raising campaigns for parents and children, and the classification of commercial content according to national standards of decency and appropriateness” (UNICEF, 2011). The framework was found to be effective by a 2010 report, highlighting that 83 mobile operators and other market operators were serving 96 per cent of EU mobile customers, implementing the Framework through codes of conduct.

In February 2009, after two years of the adoption of the code on safer use of mobile phones, a document known as “Safer Social Networking Principles for the EU” (European Commission, 2009) was launched and signed in February 2009 by 21 members of the largest social networking service operational companies across the 27 Member States. The principles focused mostly on privacy settings, including education and awareness activities and reporting practices of abuse. Staksrud and Lobe (2010) found that the compliance of services in relation to empowering users and safe use of privacy was high.

In June 2011, a “Digital Assembly Agenda” (European Commission, 2011) was promoted by the European Commission through a workshop “Every European Child Safe Online”, where the representatives of high-tech companies submitted a proposal to develop a new high-level framework of rights and responsibilities for companies and users.

In November 2011, the EU adopted the Directive of the European Parliament and the Council on combating the sexual abuse and sexual exploitation of children and child pornography and replacing Council Framework Decision 2004/68/JHA. Among other actions, the Directive aimed to:

- Criminalise forms of child sexual abuse and exploitation not currently covered by EU legislation, such as grooming, online pornographic performances and the viewing of child pornography without downloading files;
- Establish lower thresholds for applying maximum penalties;
- Ensure that offenders who are EU nationals face prosecution for crimes committed outside the EU;
- Provide child victims of the offences covered with assistance, support and protection, including for claiming compensation;
- Share data relating to the criminal convictions of sex offenders between relevant authorities in member States;

At international and national level, there are several initiatives that have shaped the environment, such as:

- The Technology Coalition: this initiative that started in the 2006 focused on the eradication of child sexual exploitation and is sponsored by the National Centre for Missing & Exploited Children (NCMEC) and the International Centre for Missing & Exploited Children (ICMEC).

- EU ICT Coalition for children Online: this initiative is built on six principles targeting young users in Europe; 20 companies are members, coming from the ICT sector.

- WePROTECT Global Alliance: a worldwide initiative concerned with child sexual exploitation, integrated by most EU Member States and others across the globe, and representatives of companies such as Microsoft, Facebook, Google on the board.

- INHOPE advisory board: INHOPE is a joint network of hotlines aimed at removing illegal content related to child sexual abuse material online. Its advisory board is made of representatives from key stakeholder organisations including Microsoft, Verizon, Google, Facebook, etc. This board provides valuable insights from the industry and an opportunity for dialogue.
- The Internet Watch Foundation (IWF): this organisation aims to minimise CSAM across the globe as well as any non-photographic CSAM that is hosted in the UK.

Key research objective: to measure how effective these incentives and capabilities have been at reducing harm.

Twitter, ISIS and other terrorist content

Social media companies have faced pressure to do more to tackle harms linked to terrorist content, such as that related to ISIS. Twitter has been extensively studied in the literature regarding the effectiveness of the platform in handling online terrorist content. Evidence (Berger, J.M, 2015) suggests that the current level of action undertaken by several platforms so far has weakened the ability of ISIS to convey information on social media. According to Conway M., et al. (2017) Twitter has been one of the key media channels used by ISIS, even before the formal establishment of the caliphate in 2014 (Conway, M. et al, 2017).

Twitter's own figures (Cheshire, T. 2017) highlight how more than 600,000 accounts promoting terrorism have been removed from the platform since August 2015. Out of these 600,00 accounts, more than half (376,890 accounts) were removed between July and the end of December 2016: mostly through automated tools. This signals an increase in terms of the efforts taken by the platform to remove such content. Conway M., et al. (2017) further investigate Twitter's efforts to tackle ISIS supporter accounts. Their results show how pro ISIS accounts have been aggressively disrupted: 65% of pro ISIS accounts in the database analysed were suspended within 70 days since being opened.

Looking at immediate moderation responses, 153 accounts[12] were detected for posting links to official ISIS content in the 24 hours between the 3rd and the 4th April 2017. Looking in more detail, 65% of these accounts were suspended within the first 17 hours of their activity, suggesting a concerted effort to impede such harmful content being distributed. Importantly, results cited in Conway M., et al. (2017) reveal that account suspension has lasting effects: even when user accounts returned online, they rarely regained the volume of friends and followers they had previously potentially hinting the role of educating users or acting as a deterrent effect. Thus, it seems that Twitter's efforts have indeed helped fragment the ability of ISIS to communicate influentially on the platform as well as reducing its ability to leverage on "throwaway"[13] accounts for disseminating ISIS links redirecting users towards content located on other platforms.

However, it should be noted that while Twitter has successfully put significant pressure to remove ISIS accounts, pressure to remove other Jihadist accounts has been less intense.

[12] This figure also includes around 50 accounts which were identified and flagged as "throwaway" accounts given that they were created on the spot without any expectations that they would stay online for significant time periods.

[13] By "throwaway" accounts we refer to accounts only created with the purpose of spreading content as much as possible before being detected and suspended

While 65% of ISIS accounts were suspended before their 70th day on the platform, less than 20% of other Jihadist accounts met the same fate (Conway M., et al., 2017)

Incentives, and economic and technical capabilities to reduce online harms

An executive survey carried out every year by Ernst & Young (ENISA, 2012) ranking the top 10 global business risk and opportunities lists social acceptance as a new and mounting demand for companies to meet ethical standards and instigated by public and user pressure.

Technical capabilities can limit the effectiveness of moderation for some types of content (Dreyfuss, E., 2017). This applies to live content which is currently unfeasible (from a technical perspective) to moderate immediately, or which would make mistaken or risk being accused of censorship. Similarly, employing large numbers of moderators is impracticable and would invariably lead to delays in the availability of content and thus encourage users to publish them anyway on different channels. Further, even using current state of the art for AI, it is impossible to automatically classify video contents which relies on context (Dreyfuss, E., 2017).

While understanding how they enable content moderation, there are economic and practical limitations of the filtering technologies which are currently rolled out to flag and remove harmful material which cannot be underestimated (Engstrom, E. & Feamster, N., 2017). A study published by the European Union Agency for Network and Information Security (ENISA, 2012) suggests that a platform's approach to tackling harms requires an economic capacity.

Key research objective: to understand the potential impact of regulation on these platforms

Regulation

Several laws and regulations in various jurisdictions apply to the online environment dating back to the early 2000s (Engstrom, E. & Feamster, 2017). Technology has however developed and continuously influences the way in which harmful and illegal content is shared (Davidson, J, 2011). This may refer to a more general reflection about regular update and evaluation of laws whether national or international tackling online safety and disruption of circles and market places in which illegal and harmful content is shared. As an example, continued evaluation of the impact (European Commission 2016b; 2017; 2018a) of the EU Code of Conduct concerning illegal hate speech (European Commission, 2016a) [14] shows that self-regulation may be a suitable option if evaluated and closely monitored. As the evaluation exercise demonstrates over four periods of evaluation at various times over three years, the number of notifications has considerably increased that were sent to the participating IT companies after each evaluation including the speed of treating the notices sent as well as speed of takedown of the content signalled (European Commission, 2019a; 2019b)[15].

The literature (Froiso, 2017) also mentions, that the introduction of the liability exemption for hosting service intermediaries in case of upload of harmful content by their users caused hosting service providers not to monitor, evaluate, or seek understanding about the of

[14] The Code of Conduct was initially signed by Facebook, Twitter, YouTube and Microsoft. During 2018, Instagram, Google+, Snapchat and Dailymotion announced the intention to join the Code of conduct.

[15] The last evaluation exercise dates back from 5 November to 14 December 2018 including 39 notifying organisations in 26 Member States. The first evaluation exercise included only 9 Member States and 12 notifying organisations. This was gradually increased with each evaluation round. Overall, the European Commission concluded that: the Code has helped to get IT companies to adopt clear Terms of Service prohibiting hate speech, 89% of flagged content is reviewed within 24 hours; IT companies provide regular training and support to reviewing staff; it has provided better partnerships between IT companies and civil society organisations; IT companies have national contact points in each Member State.

presence of harmful content or illegal activity on their domain (Ibid). Recently, policy debates emerged over introducing more secondary liability for online intermediaries[16].

Sources also highlight (Davidson, 2011) referring to the example of CSAM, that due to the inherent transnational nature of the internet, a piecemeal national regulatory approach seems ineffective regarding take down of content and prosecution of those sharing CSAM (e.g. due to lack of common definitions of what an image is depicting child abuse) or solicitating children for sexual purposes such as grooming (sometimes illegal or legal depending on the legal age to sexually consent) and impacting on how sharing of such material occurs (ECPAT International, 2018)[17]. However, international standards alone without the national political will of signatory states or a coercing international monitoring framework to enforce these standards (Carr, J., and Hilton, Z., 2011)[18], are insufficient to provide for a coherent international legal framework. It demonstrates the complexity and challenges of regulation compared to the ease of sharing harmful content within minutes across jurisdictions.

On this matter, one can refer to the UK context as an example. The IWF (Internet Watch Foundation) reported in 2017 that less than 1% of CSAM was hosted inside the UK (International Watch Foundation, 2018). The IWF highlights also in this report that from the instances that were found in the UK, hosting service providers were used as “free riders”, meaning commercial hosting services were abused to host websites containing CSAM and that in 14 out of 15 cases that were identified in 2017, the hosting service provider was not a member of the IWF (Ibid.). This could indicate that co-regulatory approaches of in the form of establishment of a hotline to tackle CSAM, alongside code of conducts for identification and takedown of material, can be an effective way to potentially reduce the hosting of such illegal material (Davidson, 2011) in one country. It however cannot prevent citizens from viewing such material online if hosted elsewhere if that approach remains national.

Technology

Intelligent technical systems are necessary to enable successful moderation of online harms at scale, given the vast amount of UGC that is uploaded to platforms by its users (De Clercq, O., et al, 2013). In the area of CSAM, the IWF’s Image Hash List (which uses PhotoDNA technology) and URL List are two systems which can be deployed by its more than 130-member organisations to moderate CSAM online, or links to CSAM content, when they appear on the platform’s services. The impact of the development of tools and technology aimed at tackling CSAM online has meant that it is increasingly difficult to use major search providers to find CSAM on the open web (WePROTECT Global Alliance, 2016). The result of this can mean that internet users looking to access or distribute CSAM online instead move to peer-to-peer (P2P) networks and use encrypted technology (Ibid.).

Despite obvious successes in certain areas, there are limitations to content filtering technology, suggesting that total reliance on technological tools for the filtering and removal of harmful content is not currently a viable option. When it comes to moderating hate speech, research

[16] The European Commission considered amending the eCommerce Directive’s liability limitations for hosting service providers and putting in place and recommending to IT companies that they introduce user-friendly notice mechanisms, have counter-notice mechanisms in place and provide for more transparency of the steps undertaken to detect and takedown content.

[17] One cannot refer to actually ‘reducing’ such type of material online as research has not yet established to what extend there is more CSAM online available or the same content changes the way of how it is shared. There is no literature reference that includes information that can demonstrate a link as to why more material may have been uncovered in one-time period compared to another period of time and why potentially CSAM is increasing or decreasing in various years.

[18] The simple adoption of UN standards was insufficient, as legal standards are not directly applicable in domestic court rooms, nor are there mechanisms in place to force signatories to be compliant with those standards, nor for other parties to take up complaints for failures under the treaties application.

has shown that its context-dependence cannot be adequately captured by various dedicated systems which have been trained to automatically detect hate speech, demonstrating a recurring inability to distinguish between offensive speech and hate speech (Engstrom, E. & Feamster, N., et al, 2017). For cyberbullying, the limited availability of public datasets that characterise cyberbullying that could be used to train classifiers that tackle the phenomenon through filtering and content removal have been cited as a 'key challenge' (De Clercq, O., et al, 2013). In addition to its various inherent limitations, filtering tools can be expensive for many start-ups or smaller platforms (Engstrom & Feamer, 2017).

To respond to these inherent limitations in technology, there is a need to develop other processes that support the technological moderation of online harms; in particular, the development of online safety materials for parents, children and adult users has been identified (HM Government, 2018).

Transparency reporting

Ensuring that platforms are transparent about the internal processes have in place to moderate USG is an important step to ensure that companies can be held to account for how effectively they attempt to address the potential negative consequences of UGC (Shipp, 2018). Further, transparency reporting by online platforms could serve to assuage the feeling of disempowerment felt by 89% of the UK population in how online products and services operate (Doteveryone, 2018). A report from 2016 found that there was a lack of transparency in general around the decisions made by online platforms behind their moderation outcomes and the ability for platform users to appeal when their content is taken down. The same report highlighted that a lack of transparency regarding the processes in place for users to recover their accounts led to user distrust in the case of Facebook's policy on 'real names' (Onlinecensorship.org, 2016).

To define how intermediaries can demonstrate best practice in their moderation of online content, efforts between civil society groups and experts have resulted in the creation of sets of key guiding principles. Among these various principles is a strong call for transparency and accountability.

The Manila Principles on intermediary liability advise that intermediaries should 'publish their content restriction policies online' in a clear and accessible format (principle 6a), updating them when necessary, and to publish transparency reports about the enforcement of their content policies and all restrictions made in response to requests from government, court orders and private companies (6e) (amongst other recommendations). Transparency reporting is also advocated by the Santa Clara Principles on Transparency and Content Moderation, which were conceived as a way of engendering accountability and transparency regarding UGC moderation and management. The principles advise that companies undertaking content moderation publish the number of posts removed and accounts which are permanently or temporarily suspended for violating policy on content (Principle 1: Numbers), provide notice about the reasons for which a users' content is removed, or for having had their account suspended (Principle 2: Notice), and to provide a 'meaningful opportunity for timely appeal' for any content removed, and account suspended (Principle 3: Appeal).

The Internet Commission, an independent initiative 'for a more transparent and accountable internet' is drafting an Accountability Framework to serve as a model to monitor the effectiveness of online content management processes. The draft framework comprises 45 qualitative and quantitative questions ordered around six areas (The Internet Commission, 2018).

Those areas are:

- Reporting: how is the platform altered to potential breaches of its rules?

- Moderation: how are decision made to take action about content?
- Notice: how are flaggers and content creators notified?
- Process of appeal: how can decision be challenges and what happens when they are?
- Resources: what human and other resources are applied to managing content?
- Governance: how are content management processes, policies and strategies overseen?

These areas build on the Santa Clara Principles by advising for the need for transparency around resource allocation to content management and internal governance structures affecting the moderation and management of UGC.

Annex 2 Research framework

Table A2.1 Research framework

Key research objectives (KROs)	Key research questions (KRQs)	Sub-questions
KRO1: To understand what different platforms define as harms on their platforms.	**1.** How do the platforms officially define an 'online harm'?	**1.1** Are there differences in definitions across online platforms?
		1.2 Do platforms differentiate between illegal and inappropriate online harms?
	2. Is there a formal mechanism in place to amend any categorisation of harms?	
KRO2: To understand the incentives, capabilities and methods of different platforms to address online harms, including technical and economic capabilities.	**3.** What are the various moderation strategies in place to tackle online harms?	
	4. Does the platform have dedicated channels in place to **report** online harms?	**4.1** Do these channels vary depending on the type of online platform?
		4.2 Does the platform provide information about review procedures of reporting/notification about online harms or abusive behaviour (moderation strategies)?
		4.3 Does the platform provide information about identification procedures of online harms or abusive behaviour?
		4.4 In what circumstances are users/ subscribers suspended from use of the service/ platform?
		4.5 Does the platform provide feedback to reporting users?
	5. Does the platform have dedicated **technological apparatus** in place to **reactively** address online harms?	
	6. Does the platform have dedicated **technological apparatus** in place to **proactively** address online harms?	
	7. What are the **incentives** for the platform to address online harms, beyond confirmed illegal content/activity?	**7.1** Do current legal regulations guide the processes in place to address online harms?
		7.2 Do current industry codes of conduct (or other self-regulatory mechanisms that the provider has publicly declared adherence to) ease the processes in place to address online harms?
		7.3 Does addressing online harms affect the competitive advantage of the company?
		7.4 Is reputational threat an incentive for the platform to address online harms?

	8. What is the annual (or otherwise) **financial resource allocation** dedicated to addressing online harms?	
	9. What is the annual (or otherwise) **human resource allocation** dedicated to addressing online harms?	**9.1** What is the ratio of Human-v-AI intervention?
	10. Do economic constraints constitute a barrier to the platform addressing online harms?	
	11. What are the external drivers which impact how platforms address online harms?	**11.1** How do legal and non-legal regulation, the work of competitors, technological change and investors act as drivers?
KRO3: To measure how effective these incentives and capabilities have been at reducing harm.	**12.** To what extent is the platform **successful in reducing** online harms?	**12.1.** Does the platform have **metrics** in place to monitor its effectiveness in addressing online harms?
		12.2. Does the platform have **metrics** in place to monitor its efficiency in addressing online harms?
	13. What are the **links** between **incentives** and **effectiveness** to reduce online harms?	
	14. What are the **links** between **economic** and **technical capabilities** and **effectiveness** to reduce harm?	
KRO4: To understand the potential impact of regulation on these platforms	**15.** What **type of (external and internal) measures have reduced** online harms or abusive behaviour on a platform most significantly?	**15.1** To what extent have technological tools contributed to proactively identifying online harms?
		15.2 Has transparency reporting improved understanding of how platforms track and remove harmful online content and of the extent to which platforms provide a safe environment for their users?
	16. What are the **economic impacts** of regulation on online platforms?	**16.1** What are costs associated with technology that proactively reduces online harms?
		16.2 What are costs associated with implementing effective reporting channels?
		16.3 What are the costs associated with transparency reporting?

Annex 3 Limitations and mitigating measures

Table 1 Study limitations and accompanying mitigating measures

	Limitation	Explanation	Mitigation measure
Accuracy and comprehensiveness of data	Limited data availability on **financial** information in desk research	Financial information related to either revenues, profits or monetary cost of moderation was perceived by respondents as highly sensitive, hence, refusing to provide any exact figure. Further, there was no publicly available data on this aspect.	Although precise or estimated figures on the monetary values of revenue streams, costs and profits are currently not available, qualitative data on the financial implications (low, moderate or high) of addressing online harms was collected.
	Limited data availability on **human resource** information in desk research	Similar to financial information, data on the staff allocated to moderation was reported only by 5 platforms.	It was challenging to estimate the cost of moderation without financial information about the value of labour dedicated to these tasks. However, whether moderation represented a burden or not was inferred for each platform that shared staff allocation. Those platforms that did not report on staff allocation were asked to comment on the financial implications.
	Limited data availability on other **commercially sensitive information relating to the business model** in desk research	Commercially sensitive data such as customer segmentation, business strategy, profit targets or unique selling points were not reported.	There was limited information to uncover the links between some elements of the business model and activities to tackle online harms. However, platform representatives conveyed views on the most important internal and external drivers affecting incentives and capabilities to address online harms.
	Self-reported information on incentives and barriers	Data on incentives and barriers were not based on factual information but on stated views from respondents, which might be biased.	There was a lack of factual and unbiased information on incentives and barriers. Hence personal observation helped gauge such information and contrast them across different platforms.
Disclosure of online platform	Need to anonymise platform by using a self-defined general descriptor in the internal version of the report	Ideally, it would have been interesting to explore patterns of behaviour by type of platform. However, reference made to any platform at a more granular level would have increased the risk of disclosing their identity. There was undoubtedly a trade-off between rich and granular analysis by platform type	To some extent, there is a gap in the links between type of platforms and behaviours related to online harms, incentives, barriers and economic capabilities. However, the fieldwork was able to look more closely at access to resources, and revenue models, which provide useful information.

	Limitation	Explanation	Mitigation measure
		and confidentiality. However, the majority of platforms would not have participated in a study of this nature they were not anonymised.	
Survey response rate	Only four platforms completed the survey, and one platform submitted a narrative response.	Granular data per harm may be missed, regarding: moderation strategies, reporting channels, financial information on costs per type of harm.	Mostly granular data per type of harm is missing, hence desk research complement this. Interviewees were able to comment on those top priority harms (e.g., terrorism, CSAM) and provide some information on moderation activities for those.
Limited source availability	Despite identifying 148 sources of relevance to the key research objectives, we found certain themes and questions (such as on the economic impacts of regulation on line harm, or whether there were links between economic and technical capabilities) were not much covered in the literature.	Given how current the issue of moderation by online platforms of online harms is, it was not surprising that some gaps were identified.	Some research questions could not be captured by the literature review; regardless, an attempt to answer them through fieldwork was attempted.

Annex 4 References

Deutscher Brundestag (2017), Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act) (2017) Available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=CC9AD0BFB6422711411608EC8C354BA4.2_cid289?__blob=publicationFile&v=2

Australian Government (2015) Enhancing Online Safety Act 2015. Available at: https://www.legislation.gov.au/Details/C2015A00024

Baines, V. (2008). *Online Child Sexual Abuse: The Law Enforcement Response*. Available at: https://www.ecpat.org/wp-content/uploads/legacy/Thematic_Paper_ICTLAW_ENG.pdf

Berger, J.M (2015), *Can We Win the War Against ISIS by Focusing on Social Media?* (online) Available at: http://www.huffingtonpost.com/jm-berger/isis-social-media_b_6733206.html

Briggs, L. (2019). *Report of the Statutory Review of the Enhancing Online Safety Act 2015 and the Review of Schedules 5 and 7 to the Broadcasting Services Act 1992 (Online Content Scheme*), report commissioned by the Australian Government [Online] available at: https://communications.govcms.gov.au/publications/report-statutory-review-enhancing-online-safety-act-2015-and-review-schedules-5-and-7-broadcasting

Carr, J., and Hilton, Z. (2011), 'Combating child abuse images on the internet – international perspectives' in *Internet Child Abuse: Current Research and Policy* Edited by Julia Davidson and Petter Gottschalk.

Cheshire, T. (2017), *Twitter removes hundreds of thousands of terror accounts*. [online] Available from: http://news.sky.com/story/twitter-removes-hundreds-of-thousands-of-terror-accounts-10809768

Conway, M., et al. (2017), *Disrupting Daesh: Measuring takedown of online terrorist material and its impacts*. (online) Available at: http://www.voxpol.eu/download/vox-pol_publication/DCUJ5528-Disrupting-DAESH-1706-WEB-v2.pdf

Counter Extremism Project (no date), *Anwar al-Awlaki's Ties to Extremists.* [online] Available at*:* https://www.counterextremism.com/anwar-al-awlaki

Cowie, H. (2013), *Cyberbullying and its impact on young people's emotional health and well-being.* Available at: https://www.cambridge.org/core/journals/the-psychiatrist/article/cyberbullying-and-its-impact-on-young-peoples-emotional-health-and-wellbeing/B7DB89A2035CF347E73D21EAF8D91214/core-reader

Crisp (2018), *4 types of content moderation every social media manager should know about.* [online] Available at: https://blog.crispthinking.com/4-kinds-of-content-moderation-every-social-media-manager-should-know

Data & Society (2018) *Alternative Influence: Broadcasting the Reactionary Rights on YouTube.* Available at:https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

Davidson, J. (2011), 'Legislation and Policy', in *Internet Child Abuse: Current Research and Policy* Edited by Julia Davidson and Petter Gottschalk.

De Clercq, O. et al. (2013), *Normalization of Dutch user-generated* content. Available at: https://biblio.ugent.be/publication/4158367

Deutscher Brundestag (2019) Kleine Anfrage der Abgeordneten Manuel Höferlin, Frank Sitta, Grigorios Aggelidis, weiterer Abgeordneter und der Fraktion der FDP (*Short Question from the Liberal Party FDP*), Reference - Drucksache 19/6739. Available at: https://kleineanfragen.de/bundestag/19/7023-evaluierung-netzdg-und-transparenzberichte

Doteveryone (2018), *People, Powre and Technology: The 2018 Digital Attitudes Report.* Available at: http://attitudes.doteveryone.org.uk/files/People%20Power%20and%20Technology%20Doteveryone%20Digital%20Attitudes%20Report%202018.pdf

Dreyfuss, E. (2017), *AI Isn't Smart Enough (Yet) to Spot Horrific Facebook Videos*. (online) Available at: https://www.wired.com/2017/04/ai-isnt-smart-enough-yet-spot-horrific-facebook-videos

Echikson, W., Knodt, O. (2018) *Germany's NetzDG: a key test for combatting online hate*. Research report, CEPS [Online] Available at: https://www.ceps.eu/publications/germany%E2%80%99s-netzdg-key-test-combatting-online-hate

ECPAT International (2018), *Trends in online child sexual abuse material*, Bangkok: ECPAT International. [online] Available at: http://www.ecpat.org/wp-content/uploads/2018/07/ECPAT-International-Report-Trends-in-Online-Child-Sexual-Abuse-Material-2018.pdf

Engstrom, E. & Feamster, N. (2017), *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools*. (online) Available at: https://torrentfreak.com/images/Engine_-_Empirical_Study.pdf and http://www.engine.is/the-limits-of-filtering

ENISA (2012), *Economics of Security: Facing the challenges: A Multidisciplinary Assessment. Publications Office of the European Union*. Available at: https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/files/EoS%20Final%20report

European Commission (2009), *Safer Social Networking Principles for the EU*. Available at: https://ec.europa.eu/digital-single-market/sites/digital-agenda/files/sn_principles.pdf

European Commission (2011), *Digital Agenda Progress Report 2011.* Available at: https://ec.europa.eu/digital-single-market/en/news/digital-agenda-progress-report-2011

European Commission (2013), *Global Alliance against Child Sexual Abuse Online. Report – December 2013.* Available at: https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/organized-crime-and-human-trafficking/global-alliance-against-child-abuse/docs/global_alliance_report_201312_en.pdf

European Commission, (2016) a, *Code of Conduct countering illegal hate speech online*. [Online] available at: https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300

European Commission (2016) b, *Code of Conduct on countering illegal hate speech online: First results on implementation*. [online] Available at: http://ec.europa.eu/newsroom/document.cfm?doc_id=40573

European Commission (2017), *Code of Conduct on countering illegal hate speech online: Second round of monitoring: results on implementation*. [online] Available at: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=71674

European Commission (2018) a, *Code of Conduct on countering illegal hate speech online: Third round of monitoring: results on implementation*. [online] Available at: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=612086

European Commission (2018) b, *European Framework for Safer Mobile Use by Younger Teenagers and Children*. [online] Available at: https://ec.europa.eu/digital-single-market/en/european-framework-safer-mobile-use-younger-teenagers-and-children

European Commission (2019) a, *Code of Conduct on countering illegal hate speech online: fourth evaluation confirms self-regulation works*. [online] Available at: https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf.

European Commission (2019) b, *How the Code of Conduct helped countering illegal hate speech online: Fact sheet.* [online] Available at: https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf.

Europol (2015), *Strategic Assessment of Commercial Sexual Exploitation of Children Online: European Financial Coalition Report*. Available at: https://www.europol.europa.eu/publications-documents/commercial-sexual-exploitation-of-children-online

Facebook (2018) NetzDG Transparenzbericht [Online] Accessed at: https://fbnewsroomus.files.wordpress.com/2018/07/facebook_netzdg_juli_2018_deutsch-1.pdf p. 2

Frosio, G. 2017. *Reforming Intermediary Liability in the Platform Economy: a European Digital Single Market Strategy.* Available at: https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1250&context=nulr_online

Haynes, D., Bawden, D. and Robinson, L. (2016). *A regulatory model for personal data on social networking services in the UK.* Available at: http://openaccess.city.ac.uk/14916/8/Regulatory%20Model%20-%20manuscript%20v1%201.pdf

HM Government (2018), *Government response to the Internet Safety Strategy Green Paper.* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf

Internet Watch Foundation (2018), *Annual Report*. Available at: https://annualreport.iwf.org.uk

John A., Glendenning A.C., Marchant A., Montgomery P., Stewart A., Wood S., Lloyd K., Hawton K., (2018), *Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Revie*w. Available at: https://www.jmir.org/2018/4/e129/

LOI n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1) (*Law relating to the fight against manipulation of information*), published in Official Journal JORF n°0297 du 23 décembre 2018, texte n° 2, accessed at : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037847559&dateTexte=&categorieLien=id

Lynskey, O. (2017), *Regulating 'Platform Power'.* Available at: http://eprints.lse.ac.uk/73404/1/WPS2017-01_Lynskey.pdf

Mills, R. (2013), *Distributed moderation systems: an exploration of their utility and the social implications of their widespread adoption.* Available at: http://eprints.lancs.ac.uk/70128/

Office of Parliamentary Counsel, Canberra (1992) Broadcasting Services Act 1992, last amended 18 September 2018, available at: https://www.legislation.gov.au/Series/C2004A04401; Schedule 5 of the Act concerning online services, accessed at: http://www6.austlii.edu.au/cgi-bin/viewdoc/au/legis/cth/consol_act/bsa1992214/sch5.html; Schedule 7 of the Act Content Services, accessed at: http://www6.austlii.edu.au/cgi-bin/viewdoc/au/legis/cth/consol_act/bsa1992214/sch7.html

Onlinecensorship.org (2016), *Unfriending censorship: Insights from four months of crowdsourced data on social media censorship*. Available at: https://s3-us-west-1.amazonaws.com/onlinecensorship/posts/pdfs/000/000/044/original/Onlinecensorship.org_Report_-_31_March_2016.pdf?1459436925

Parliament of Australia (2018) a, *Enhancing Online Safety (Non-consensual Sharing of Intimate Images) Bill 2018, A Bill for an Act to amend the Enhancing Online Safety Act 2015, and for other purposes*. Available at: https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/s1113_third-senate/toc_pdf/1727820.pdf;fileType=application/pdf

Parliament of Australia (2018) b, *Telecommunications and Other Legislation Amendment (Assistance and Access) Bill (2018),* available at: https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r6195

Rees., M. (2018), *La proposition de loi contre les fausses informations suscite l'appétit des deputes*, [Online News Portal], article published 21 mai 2019, available at : https://www.nextinpact.com/news/106615-la-proposition-loi-contre-fausses-informations-suscite-appetit-deputes.htm?skipua=1

Reporter ohne Grenzen (2018), *NetzDG führt offenbar zu Overblocking*. Press release [Online] Available at: https://www.reporter-ohne-grenzen.de/pressemitteilungen/meldung/netzdg-fuehrt-offenbar-zu-overblocking/

Sénat (2018) a, *Lutte contre la manipulation de l'information (PPLO).* [online] Available at: http://www.senat.fr/enseance/2018-2019/29/Amdt_1.html

Sénat (2019) b, *Comptes rendus de la commission de la culture, de l'education et de la communication* [online] available at: http://www.senat.fr/compte-rendu-commissions/20180402/cult.html

Shipp, J. *A more transparent and accountable Internet? Here's how.* (online) Available at: https://blogs.lse.ac.uk/mediapolicyproject/2018/05/24/a-more-transparent-and-accountable-internet-heres-how/

Social Media Today (2010), *6 types of content moderation you need to know about.* [online] Available at: https://www.socialmediatoday.com/content/6-types-content-moderation-you-need-know-about

Staksrud, E. and Lobe, B. (2010*) Evaluation of the implementation of the Safer Social Networking Principles for the EU Part I: General Report.* Available at: https://www.duo.uio.no/bitstream/handle/10852/27216/Safer-Social-Networking-part1.pdf

UK Home Office (2017) *Home Office to crack down on online child sexual abuse with new cutting-edge technology*., [Online] accessed at: https://www.gov.uk/government/news/home-office-to-crack-down-on-online-child-sexual-abuse-with-new-cutting-edge-technology

Unicef, 2011. Child safety online: Global challenges and strategies. UNICEF Innocenti Research Centre. Available at: https://www.unicef-irc.org/publications/650-child-safety-online-global-challenges-and-strategies.html

United Nations Office on Drugs and Crime (2012), *The use of internet for terrorist purposes.* Available at: https://www.unodc.org/documents/frontpage/Use_of_Internet_for_Terrorist_Purposes.pdf

Wales, T. (2009), *Industry self-regulation and proposals for action against unlawful filesharing in the UK: reflections on Digital Britain and the Digital Economy Bill.* Available at: https://www.oii.ox.ac.uk/archive/downloads/publications/IB5.pdf

WePROTECT Global Alliance (2016), *The WePROTECT Global Alliance. Our Strategy to End the Sexual Exploitation of Children Online'.* Available at: https://static1.squarespace.com/static/5630f48de4b00a75476ecf0a/t/578408b5f7e0ab851b789e14/1479254482761/WePROTECT+Global+Alliance+Strategy.pdf

Wolak, J. et al. (2018), *Sextortion of Minors: Characteristics and Dynamics*. Available at: https://www.jahonline.org/article/S1054-139X(17)30423-8/fulltext

Annex 5 Expert Advisory Board

We recruited the support of an Expert Advisory Board (EAB) to support us at strategic moments throughout the study. The EAB comprised Dr Victoria Baines, Professor Ian Walden and Michael Moran.

Dr Victoria Baines is a researcher with a professional background in criminal intelligence analysis (CEOP and Europol), and recent experience in trust and safety for a global technology company. She conducts research on behalf of academic institutions and civil society organisations in the fields of child protection, cybercrime, surveillance and digital ethics. She is a trustee of the Lucy Faithfull Foundation, and a member of the INHOPE Advisory Board

Professor Ian Walden is Professor of Information and Communications Law at the Centre for Commercial Law Studies, Queen Mary University of London and Of Counsel to Baker McKenzie.

Michael Moran is Coordinator for the Crimes Against Children Department and Sub-Director of the Trafficking in Human Beings Department at Interpol, a position he has held since 2006. In this role, he deals with online child exploitation investigation.