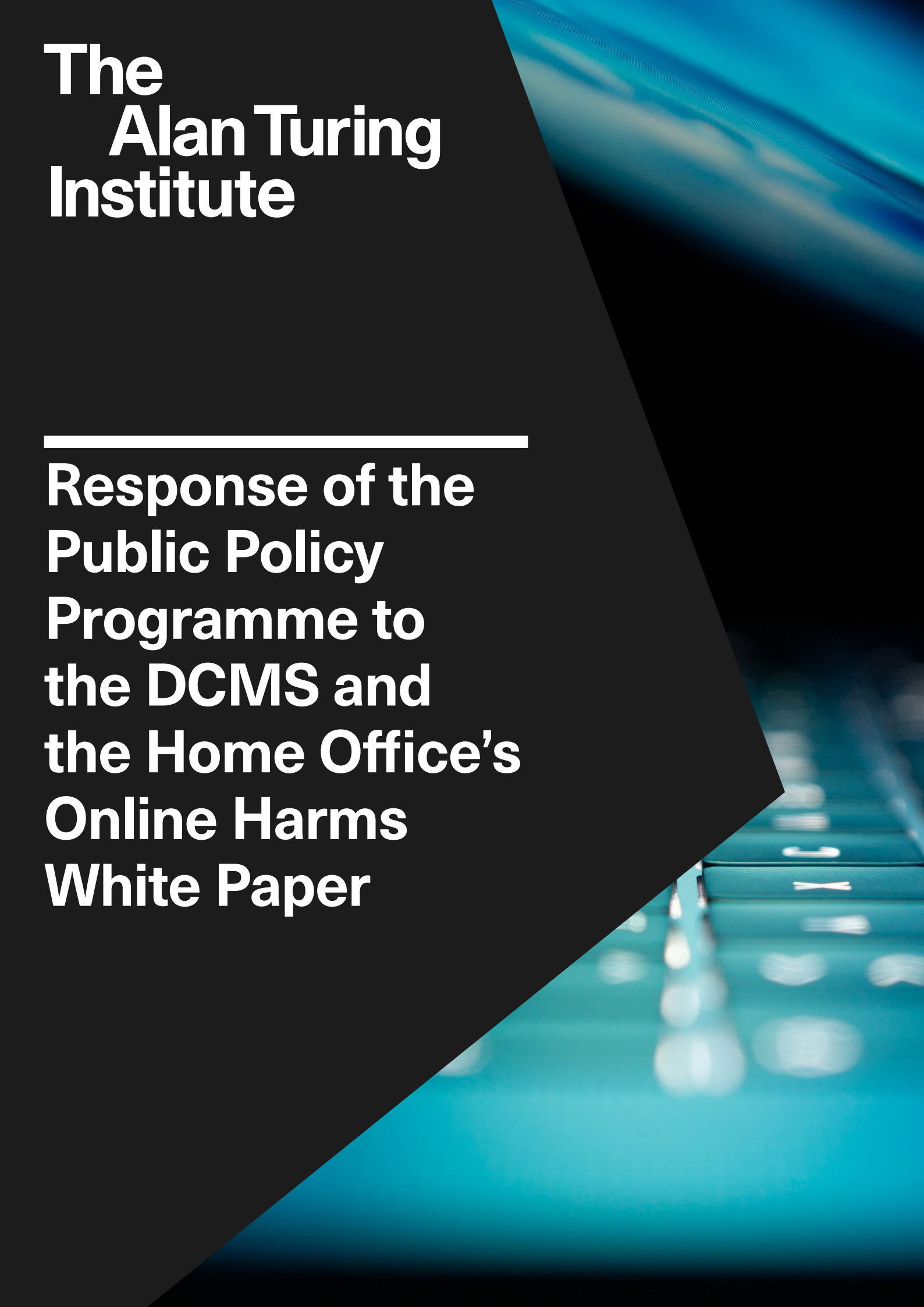
The Alan Turing Institute

Response of the Public Policy Programme to the DCMS and the Home Office's Online Harms White Paper



Response of The Alan Turing Institute's Public Policy Programme to the Department for Digital, Culture, Media & Sport and the Home Office's Online Harms White Paper

Introduction

This document provides the response of The Alan Turing Institute's Public Policy Programme to the Department of Digital, Culture, Media & Sport (DCMS) and the Home Office's Online Harms White Paper (April 2019). The response represents the views of the Public Policy Programme. A list of researchers who contributed to the response can be found in the Appendix.

We congratulate DCMS and the Home Office on producing this White Paper. It marks an important step forward in achieving better regulation of the Internet and shows the UK's commitment to being at the forefront of responsible Internet governance. The broad message of the White Paper is commendable: "We cannot allow these harmful behaviours and content to undermine the significant benefits that the digital revolution can offer [...] If we surrender our online spaces to those who spread hate, abuse, fear and vitriolic content, then we will all lose." (p.3) The cornerstone of the White Paper is establishing that digital platforms have a statutory duty of care for individuals who use their services. This is a positive step forward in achieving safe and responsible governance of the Internet.

This is an extensive White Paper with a broad scope. However, we feel several issues are left unresolved. For example, we would like a high-level explanation of what constitutes a harm, and how differing harms will be prioritised. This is a complex area, with many challenges and obstacles still to be overcome in order to provide effective, clear and world pioneering governance and regulation of online spaces. We commend the steps this White Paper has taken while acknowledging the work still required to meet these challenges.

The Alan Turing Institute's Public Policy Programme is open to engaging further in the development of a regulatory framework for online harm and welcomes any questions regarding this response. The wider Institute is referenced in two parts of the White Paper, both on p. 80: (1) the creation of the Digital Charter Fellowship programme, which is a joint initiative between DCMS and the Turing and (2) the Turing-led research project *Hate Speech: Measures and Counter Measures,* which develops techniques for monitoring and tackling online hate speech.[1]

[1] The Alan Turing Institute, 'Hate Speech: Measures and Counter-Measures', Available at: https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures, Last accessed on: 20th June 2019.

This response is structured in two sections. In Section 1, we consider important issues which are not directly addressed by the Consultation questions. In Section 2, we provide responses to 8 of the 18 Consultation questions.

Section 1 | Specific issues

The regulator

The White Paper advocates creating a new independent regulator. However, it also notes several issues with this and raises the possibility of using an existing regulatory body. One challenge for a new regulator is that online harms are multifaceted and varied, spanning many policy domains. We feel a single regulator might struggle to develop enough internal expertise to tackle them all at once and to identify and respond to new risks as they emerge. There could be communication and co-ordination problems with too many separate government bodies operating in this space, resulting in a lack of clarity for both citizens and the digital tech platforms. The broad remit of the regulator means that it will likely have to operate across many areas of Government, including the Police, Security Services, Department of Education, Ministry of Justice, DCMS and the Home Office, as well as navigating the complementary responsibilities of the Office of Communications (Ofcom) and the Information Commissioner's Office (ICO). Meanwhile, existing regulators have accumulated much of the expertise needed in dealing with data-intensive digital platforms that the regulation of online harms will require. For these reasons, we recommend that a new unit with a specific remit for online harms be established within one of the existing regulators, such as Ofcom or the ICO.

Cross-platform response

At present, we believe that the White Paper does not place enough emphasis on cross-platform measures for tackling harmful content. Two important aspects of online behaviour should be explicitly considered.

First, is how users migrate between platforms. At present, digital platforms are free to formulate their own approach for defining, detecting and moderating harmful content, without any requirement or incentive to work with other digital platforms. This is likely to have unintended negative consequences. For example, if a mainstream platform bans a prominent figure engaged in hate speech from their platform, then they may reduce the spread of this harmful content. However, that individual may join a smaller niche platform in which there is little or no content moderation. They may also encourage some of their supporters to migrate with them. These supporters could then be at greater risk of extremism as they engage in more hateful discussions, free from mainstream content moderation and dissenting voices. Thus, whilst a mainstream platform is 'cleaned up', it could be at the expense of some people

engaging in more extreme spaces. Whether this is, overall, a positive or negative development is up for discussion – and whether pre-emptive actions can be taken to avoid these negative outcomes requires further investigation.

Second, is how harmful content moves between different platforms. Research suggests that content moves from niche extremist platforms, such as 4chan, to big mainstream platforms, such as Twitter and Facebook (Hine et al. 2016). The White Paper does not contain any explicit discussion of this, and how platforms are inter-related. This is especially important for developing strategies to understand (and tackle) harmful behaviour within the wider online ecosystem.

We believe both these challenges should also be addressed by the unit tasked with regulating online harms as, by their very nature, they transcend the powers and remit of any single platform. We propose that the major benefits of a regulatory unit which works across the industry is that they can (1) develop a deep understanding of these challenges through evidence-gathering and original research and (2) develop a cross-platform joined-up regulatory approach to content moderation.

We believe that the White Paper should have closer regard to the dynamics of online harms, and how they span and migrate between multiple, changing platforms.

Truth and facts

The White Paper states several times, 'We are clear that the regulator will not be responsible for policing truth and accuracy online.' (p. 36). However, we are unsure how disinformation (discussed in detail on pp. 22-24) will be addressed without the regulatory unit either taking a position on truth/falsity of content or mandating another body (such as the platforms or a 3rd party) to do so. We would also welcome greater clarity on what is 'in scope' of the White Paper's understanding of misinformation as the range of different types of false or misleading information is considerable, including (1) explicitly false content, (2) misleading interpretations of facts, (3) partial and one-sided analyses, and (4) predictions and opinions treated as facts. More broadly, following on from the White Paper, we hope the roadmap will describe a regulatory response beyond describing the problem of misinformation. Even though disinformation does not form part of the online harms singled out for proactive monitoring provisions, we feel some form of definition is still required to assess the effectiveness of 'softer' tactics, such as flagging or contextualising content. We would welcome more detail to better understand how (and if) the regulatory unit will act and how this will impact freedom of expression, independence of the media and civic discourse.

We feel the White Paper should be more consistent on whether and how disinformation is within scope, and which bodies and criteria (if any) will be used to assess what is true or false.

Freedom of expression

We would welcome a policy roadmap after the White Paper consultation that provides a discussion on how to balance freedom of expression with the need to protect individuals from harm.

We welcome the gravity of the Government's concerns regarding restrictions on freedom of expression. We believe that freedom of speech is an important issue and should always be protected. However, we would like to see further discussion of how freedom of speech should be balanced against the need to protect individuals from harm in a future policy roadmap. This is a complex and difficult issue, which will require a substantive discussion of what values should be incorporated into the regulatory framework, with potentially contentious decisions needing to be made regarding constraints on online behaviour.

The White Paper mentions that the regulator must protect freedom of expression, and we would welcome additional discussion on how this should happen. Protecting the rights of users is covered in only one paragraph (point 5.12). We would also welcome further detail on how users will be protected against potentially overzealous automated take-down systems. Machine learning systems remain inconsistent at understanding the broader context of a piece of content, such as its role in an ongoing social interaction, or whether it was said in jest or as a parody. They may also be biased and discriminatory on the basis of the populations on which they were trained, or misclassify minority dialects (Binns et al., 2017; Badjatiya et al., 2019; Blodgett and O'Connor 2017). Furthermore, many pieces of content which might seem harmful may actually be in the public interest, such as videos of war atrocities aimed to draw attention to breaches of international law. For these reasons, we feel **it is important that users' ability to contest any use of takedown technology, even if used in a voluntary manner, is also subject to regulatory oversight.**

Evidence and research

Evaluating the prevalence of harmful online behaviour is difficult due to a lack of appropriate definitions, measurements and data. As such, while it is likely that 'online harms are widespread' (as stated in the ministerial foreword, p. 3, p. 5, and p. 12), we think far more research is required to ensure this is correct. More broadly, we hope that this White Paper acts as a catalyst for more investment into research on online harms, as a better evidence base is urgently needed. This will require deep collaboration between industry, government, academia and civil society. This is noted in the paper (point 43, p. 9), and we would welcome plans for up-scaling existing collaborations (such as wider partnerships and increased funding).

A stated goal of the new regulator is to encourage cross-partnership collaboration (point 24 on p. 8). We would welcome more detail as to how this will work in practice. We would also

welcome actions taken by the regulatory unit to encourage digital platforms to share valuable resources, such as datasets and code, as well as to provide greater transparency about their content moderation policies and processes. At the same time, we accept that there needs to be a degree of confidentiality, especially given the risk of rogue users 'gaming' moderation systems. Some of the large digital platforms, such as Facebook and Twitter, have sought to make datasets and funding available for research, which we strongly encourage.

Additionally, in some cases, we believe that the Internet is often blamed for social problems which may not be related to online activity. For instance, point 1.14 (p. 13) describes the level of knife crime and homicides in the UK. Whether this is related to online activity is not well-established, and we would encourage further research to be undertaken to establish a link prior to any future references to online activity and knife crime.

System design

The White Paper draws attention to the need to improve both how platforms moderate specific bits of content and how the digital ecosystem is regulated overall. We welcome this, as it is important that the design and functionality of platforms is addressed by the regulatory unit. In this regard, we note research by Turing academics which indicates the cognitive biases which drive uptake of online content and should be accounted for in platform design (Hills 2018). However, we would like additional clarity about what parts of digital platforms, and the broader digital ecosystem, are considered problematic and why. **Without a clearer understanding of how system design impacts user behaviour, and which aspects are considered undesirable, we think it would be difficult to see how the discussions in the White Paper will translate into policy in a desirable timeframe.** Furthermore, while we welcome greater transparency in how digital platforms are designed (especially the use of algorithms in content feeds, such as the curation of 'recommended' content), we are unsure of the need for the Government to mandate the design of platforms' systems. Overall, focusing on how platforms and the broader digital ecosystem are designed is positive, and we would welcome more detail.

Worker welfare

The White Paper focuses on the impact of online harms on users. In addition, we would welcome an explicit consideration of the impact of online harms on those that regulate the environment for users, conduct legal proceedings against criminally harmful behaviour and research online harms. These individuals include platform employees, civil society activists and leaders (such as advocacy charities), academic researchers, lawyers and journalists. We think more attention needs to be paid to the emotional and mental health challenges faced by anyone coming into regular contact with harmful content, as shown in the BBC Four *Storyville*

Documentary, "The Internet's Dirtiest Secrets: The Cleaners"[2], and articles such as "The Trauma Floor".[3]

Recent research from the Turing's *Hate Speech* project highlights the lack of robust support processes in place for researchers investigating online abuse (Vidgen et al. 2019). This publication provides a checklist of actions that individuals could use to potentially reduce and mitigate the impact of being exposed to harmful content.[4] There is also a risk that anyone working to regulate, moderate, research or counter online harms may become a victim of abuse through targeted attacks. These range from the propagation of false information about a person to 'doxing' (the publishing of identifying features online such as phone number) and 'swatting' (reporting a false threat to the police, which is usually high-level enough to warrant an armed response). We welcome further discussion and consideration of employee and researcher ethics in relation to online harms. This should include establishment of best practices and appropriate guidelines so that those who are regularly exposed to harmful content can do so in the safest and most supportive environments possible.

[2]BBC, 'The Cleaners', Available at: https://www.bbc.co.uk/programmes/m0003f2f, Last accessed on: 20th June 2019.

[3] The Verge, 'The Secret lives of Facebook content moderators in America', Available at: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona, Last accessed on: 20th June 2019.

[4] Vidgen et al., 'Challenges and frontiers in abusive content detection, Appendix', Available at: https://github.com/bvidgen/Challenges-and-frontiers-in-abusive-content-detection, Last accessed on: 20th June 2019.

Section 2 | Responses to Questions

Question 1: This government has committed to annual transparency reporting. Beyond the measures set out in this White Paper, should the government do more to build a culture of transparency, trust and accountability across industry and, if so, what?

We think digital platforms should make information about their activities to counter online harms available to researchers and civil society. This would open them up to scrutiny and enable 'shadow reporting', which is a widely practiced in the sphere of human rights law, where civil society organisations submit independent reports about the state of human rights within different jurisdictions. Experience and procedures from this field can be used to identify best practices. Such an open approach would also create synergies with the White Paper's goal of providing more information and data to researchers. The Commission for Countering Extremism recently showed best practice in this regard, releasing a global call for partnerships with researchers across a variety of topics.

The Government should involve and fund outside bodies (including academic institutions) to develop innovative and practical proposals for improving transparency, trust and accountability. An initial focus should be on providing robust definitions of each of these terms.

Question 5: Are proposals for the online platforms and services in scope of the regulatory framework a suitable basis for an effective and proportionate approach?

We think the scope of platforms and services covered by the regulatory framework appears very broad. It covers any online space where user-generated content can be uploaded, found, or where people can interact with each other. It also appears to cover designers of tools, which may be open-source, deployed in a decentralised manner, and used by a range of different communities. We would welcome more clarity on how companies will be considered in and out of scope. The White Paper states, 'The regulator will take a risk-based approach, prioritising action to tackle activity or content where there is the greatest evidence or threat of harm, or where children or other vulnerable users are at risk.' (point 34, p. 9). However, as we discuss below in response to Question 8, we believe the definition and prioritisation of harms requires additional scoping. As such, we would welcome greater clarity on which online spaces will come under the purview of the regulatory unit.

We would also welcome more clarity around how the codes of practice will be adapted to different platforms as we feel there are considerable differences between large and small platforms. While the digital environment is dominated by the big platforms, such as YouTube,

Twitter and Facebook, harmful behaviour often occurs on small niche platforms. Given different corporate cultures, internal resources and user bases, we feel a 'one-size fits all' approach would not be as effective as a more nuanced approach. Specifically, we would like to see detailed examples of how smaller 'high risk' platforms would be regulated, and what the different regulatory challenges they pose are. We would also like to see an initial list of in-scope platforms.

Another scoping issue concerns harms that can be associated with platforms and services that interact with ambient physical environments. Considerable recent discussion has centred around the use of, for example, home assistants in both detecting harms (e.g. being used as evidence in court) and in being tools of abuse themselves, such as in the context of intimate partner violence (Freed et al., 2017; Freed et al. 2019; Levy 2019). We believe the White Paper would need to specify the extent to which these ambient systems are or are not in scope. If they are in scope, we think that the White Paper should be specific about the extent to which, if at all, proactive monitoring obligations would apply. If so, we believe this would pose considerable challenges, as monitoring an online forum is very different to monitoring audio in an individual's home.

The newness of the proposed regulation makes it difficult for us to assess whether it will be effective and proportionate and we welcome additional information on implementation plans. We are also conscious of the fast-changing nature of online behaviour, and the limited evidence in this area, which would likely complicate any implementation plans. We think there is considerable potential for unforeseen negative consequences. For instance, fear of high compliance costs and punitive regulation could drive platforms to severely constrain the functionality they offer users, such as by limiting messenger services, which would be detrimental to the idea of a free and open Internet. Regulation that impacts platforms which also function as publishers, such as Wikipedia, is also likely to impact on controversial international freedom of expression issues. It may also disproportionately affect smaller platforms and innovative software developers. As such, we would encourage the Government to allocate resources to study the initial impacts of regulation before setting its final scope.

Question 6: In developing a definition for private communications, what criteria should be considered?

The White Paper recognizes that while most digital platforms are privately run and owned, they 'have become akin to public spaces' (p. 6). Nonetheless, protecting the privacy of users is crucial when monitoring harmful online content; as the White Paper states, 'the importance of privacy, any requirements to scan or monitor content for tightly defined categories of illegal content will not apply to private channels.' (point 33, p. 8). Previous research suggests that in many contexts, users are not aware that their activity is playing out in a public space, viewing their personalised social media as a quasi-private space (Marwick and Boyd 2010). We believe

further research is needed to understand how users view different spaces and their privacy rights within them, especially following the Cambridge Analytica scandal.

Research at the Turing shows that people highly value their data and seek to protect it (Skatova et al. 2019). At the same time, not all data is valued equally: people value financial data and medical records highly but are less concerned about privacy in other forms of data, such as loyalty cards, physical activity or energy use (Skatova et al. 2019). We believe this needs to be carefully considered when determining the scope of actions to tackle and remove harmful content, and the amount of personal data that is monitored, analysed and stored. Initially, what constitutes private communication can be considered by evaluating factors which influence expectations of privacy. For example, these include:

- The number of users with access to the space.
- The barriers to access and sharing. For example, if a discussion takes place in a forum with no barriers to access, such as a comments section on a news website, it would likely be considered public.
- The goal and purpose of the space. For example, if a Facebook Page is used to provide cinema listings, then it might be perceived to be more 'public' than if it is used to organise political activity.
- Whether platforms provide privacy preserving technology, such as encryption. This is an important consideration. However, as the general public may not be aware of all the details around such technologies, we suggest this should not be used as the sole determinant of 'privacy'. We believe people who communicate on non-encrypted chat services might still expect their communication to be private.

Question 8: **What further steps could be taken to ensure the regulator will act in a targeted and proportionate manner?**

To act in a targeted and proportionate manner, we believe that the regulatory unit will need to first determine what constitutes a harm and how harms can be prioritised. We have several recommendations.

First, we think the discussion of harms requires additional nuance and clarity. For instance, we feel there should be greater emphasis on the degree of harm, how 'risks' should be defined and quantified, and how the impact of harms should be assessed. The new regulatory unit, in particular, will need to consider how the scale and severity of harms vary.

Second, we would welcome greater detail around which harms are considered in-scope. We believe it would be beneficial to have a set of coherent criteria outlining *what* is considered a harm and why the identified harms should come under the purview of the regulatory unit.

Third, we would welcome a set of robust criteria for how the 'harmful but legal' category is defined. While we welcome this category, greater detail is required as regulating legal (albeit harmful) behaviour is likely to be contentious. Greater clarity on whether this is proposed as an exclusive category (e.g. some harms, such as misinformation, are only harmful but legal) or whether it is a category that can be applied to any harm (e.g. hate speech might sometimes be illegal but other times 'harmful but legal') would be helpful. Without sufficient nuance and clarity in a statute based on the White Paper, we believe there is a risk that it will not comply with existing legislature. We think that the harms discussed must be sufficiently well-defined that a platform can, before engaging in any design or moderation decisions, know the potential legal consequences that would follow.

In addition, the White Paper discusses 'unacceptable' behaviour (p.11). We would encourage the Government to clarify the definition of unacceptable behaviour and explain whether it is the same as the 'harmful but legal' category. We imagine explaining the difference in definitions might be challenging, so consideration should be given to its future inclusion in future iterations of the policy.

Fourth, the White Paper recognises that some harms are less clearly defined (p. 31). However, we suggest that some of the harms which have been described as having a 'clear definition' require additional nuance as they are complex and messy phenomena, such as harassment, hate crime, terrorist content and modern slavery. We would welcome clear definitions for each of the harms that will be regulated.

Fifth, we are unclear how the distinction between 'clear definition' and 'less clear definition' has itself been determined. We would welcome more clarity on whether this distinction is based on existing legal frameworks, academic input or other criteria.

Sixth, the harms which are listed can manifest in very different ways. For instance, online harassment can, among other possibilities, (i) actually take place online (whereby one user harasses another), (ii) be planned online but conducted offline (e.g. when groups of people discuss how they will target another person) or (iii) happen offline but reported online (e.g. by sharing and glorifying a previous activity). We would encourage the Government to provide detail around how, for different harms, online spaces are used in different ways.

Seventh, the White Paper emphasises 'minimising undue burdens' on the digital platforms (point 4.5, p. 49). We think this could be misinterpreted in how 'proportionality' is determined and result in an approach to regulation and enforcement that might be too soft. We would welcome additional clarity on these burdens, and their role in protecting online users from harm, before there is any stipulation on minimising them.

Overall, we believe that greater clarity is needed in how harms are defined and prioritised. Providing this clarity will only become more important as the digital landscape changes and

new harms emerge and will allow for a more precise assessment of how the regulatory unit will respond in a proportionate and targeted manner. We advise that the regulatory unit work in close collaboration with other public bodies active within the field of digital technologies, such as ICO and the Centre for Data Ethics and Innovation (CDEI) to determine what is considered targeted and proportionate.

Question 12: Should the regulator be empowered to i) disrupt business activities, or ii) undertake ISP blocking, or iii) implement a regime for senior management liability? What, if any, further powers should be available to the regulator?

We believe that the regulatory unit should have the power to impose sanctions on non-compliant companies. This should, however, be considered in relation to the scale of businesses: smaller companies and non-profits, for example, could be deterred by these measures, which could damage the UK technology sector. ISP blocking is a powerful tool and should be used carefully, as the Turing project *Automated discrimination in internet filtering* shows.[5] If ISP blocking is considered, we think it should be implemented in a targeted and proportionate manner and, where possible, access should be limited only to those pages with harmful content rather than to entire platforms. Disrupting platforms' operations should be balanced against the interests of the people who rely on the platforms' services. We think fines will likely be an important part of the regulator's arsenal and should be set at a sufficiently high level so as to deter platforms from enabling/allowing harmful behaviours.

Question 15: What are the greatest opportunities and barriers for (i) innovation and (ii) adoption of safety technologies by UK organisations, and what role should government play in addressing these?

We would welcome greater clarity on the use of the terms 'innovation' and 'safety' (p. 56, point 5.11) in the context of the White Paper. For example, how should the merely new be distinguished from the truly innovative? Or how safe is safe enough? We feel an open discussion on the meaning of these terms would help refine them and ensure their widespread acceptance.

The White Paper discusses the use of technology to realise its regulatory goals. We firmly welcome this as the only scalable way of tackling online harms and as a way of reducing the psychological pressure and stress on content moderators. However, we think computation is no panacea and employing more advanced methods to address these difficult problems will not necessarily solve them. We believe that computational methods do not remove the need

[5] The Alan Turing Institute, 'Automated discrimination in Internet filtering', Available at: https://www.turing.ac.uk/research/research-projects/automated-discrimination-internet-filtering, Last accessed on 24th June 2019.

for defining tasks and interrogating the social aspects of issues. Indeed, we think they are most effective when used for well-defined tasks and integrated with social science and qualitative insights.

Greater discussion of the ethical and social risks of using technology and the potential unfairness it (re)produces is needed. For instance, research suggests that most hate speech detection systems are more effective at detecting hate against certain targets, and from certain individuals, than others (Binns et al., 2018). We think this may unfairly allow certain users to systematically avoid penalties for sharing hate, while other users are systematically given less protection. The way that data-driven tools are developed and used can greatly affect their outcomes and their impact on people. We believe it is important that ethics are carefully considered by the regulatory unit, especially given the implications of using algorithmic tools for content moderation.

To help mitigate the concerns outlined here, we suggest that (1) users are given a means of recourse. Users should be able to quickly and easily assert their 'right to challenge' when their content has been removed from a platform, especially as accuracy of content moderation systems are currently underwhelming. (2) The regulatory unit has a statutory responsibility for ensuring that these rights are effectively provided. We think there should be clear accountability for moderating content, assessing claims and (if required) re-allowing content. All information should be provided to users in an accessible format.

Question 17: Should the government be doing more to help people manage their own and their children's online safety and, if so, what?

We think there needs to be a greater understanding regarding how children use their connected devices and how other actors influence this, such as the role played by (elder) siblings, the family setting, and the context of playmates and friends. The focus on digital literacy is welcome, and we believe that additional provisions are required to ensure that users are fully protected from online harms.

The ICO recently published a consultation document, titled *Age appropriate design: a code of practice for online services*[6]. The Alan Turing Institute offered a response, which is available online.[7] As part of this, we highlighted how children may circumvent restrictions and regulations. They are increasingly tech-savvy and operate across multiple devices. They may

[6] ICO, 'Age appropriate design: a code of practice for online service', Available at: https://ico.org.uk/media/about-the-ico/consultations/2614762/age-appropriate-design-code-for-public-consultation.pdf, Last accessed on 24th June 2019.

[7] The Alan Turing Institute, 'Response to the Information Commissioner's Office consultation on its Age Appropriate Design Code', Available at: https://www.turing.ac.uk/research/publications/information-commissioners-office-consultation-its-age-appropriate-design-code, Last accessed on 24th June 2019.

find parental oversight too restrictive and will want to use the platforms where their friends are active. As such, they might use fake birthdays to overcome age verification, manually change data settings to avoid restrictions and constraints without informing their parents, and move to unregulated niche platforms. If code standards are not agreed, developed, and then implemented to take these factors into account, then there is a risk that they will not be effective. We also firmly encourage the Government to commit further funding to initiatives around digital literacy targeted at children.

Question 18: What, if any, role should the regulator have in relation to education and awareness activity?

We believe that education and awareness are often unsung activities and can be easily dismissed as soft regulatory instruments. We think they are essential for countering online harms and should be amply funded. We would encourage the regulatory unit to develop schemes with outside bodies, such as relevant education institutions, departments, and civil society organisations, to enable further education and awareness-raising activities. Some of the topics that should be covered with both children and adults include:

- Sexualisation online. The White Paper refers to how underage individuals use digital tools to share sexual images of themselves. It is important that pre-teens and teenagers are well-aware of the dangers of this, which should tie into education regarding the role that sex and intimacy play in a healthy life.
- Skills to debate, reason, and empathise online. There are concerns that in online contexts, anonymity and perceived distance between persons has led to individuals losing the ability to have sincere, empathetic, in-depth and well-reasoned discussions. We think the proliferation of trolls online is partly because the human consequences of individuals' behaviour is not perceived behind a screen. We think training to increase respect, curiosity and empathy should be developed for all citizens to help mitigate this.
- Skills to critically analyse information, cross-check facts from multiple sources and assess the reliability of information. We think these are increasingly necessary as it becomes difficult to separate fake news from the truth. In the future, we fear fake news will likely gain new strength as 'deep fakes' become more widespread.

Bibliography

Badjatiya, P. et al (2019) 'Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-Based Generalizations', *The World Wide Web Conference*.

Binns,R. et al. (2017) 'Like Trainer, like Bot? Inheritance of Bias in Algorithmic Content Moderation' in Giovanni Luca Ciampaglia, Afra Mashhadi and Taha Yasseri (eds), *Social Informatics: 9th International Conference*.

Blodgett, S. & O'Connor, B. (2017) 'Racial disparity in natural language processing: a case study of social media African-American English', *ArXiv:1707.00061v1* .

Freed, D. et al. (2017) 'Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders' Proceedings of ACM Human-Computer Interactions, 1:1, pp. 1-20.

Freed, D. et al. (2019) 'A Stalker's Paradise: How Intimate Partner Abusers Exploit Technology', *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Hills, T. (2018), 'The dark side of information proliferation', *Perspectives on Psychological Science*, 14:3, pp. 323-330.

Levy, K. (2015) 'Intimate Surveillance', SSRN Scholarly Paper ID 2834354.

Marwick, A. & Boyd, D. (2010), 'I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience', *New Media and Society*, 13: 1, pp. 114-133.

Skatova, A. et al. (2019) 'Unpacking privacy: Willingness to pay to protect personal data, Available at https://psyarxiv.com/ahwe4/ .

Vidgen, B. et al (2019) 'Challenges and frontiers in abusive content detection', *Forthcoming*.

Appendix

The following researchers contributed to this response:

Alex Harris (Research Assistant, The Alan Turing Institute)

Anya Skatova (Turing Fellow, University of Bristol)

Bertie Vidgen (Research Associate, The Alan Turing Institute)

Charles Raab (Turing Fellow, University of Edinburgh)

Christina Hitrova (Research Assistant, The Alan Turing Institute)

Cosmina Dorobantu (Deputy Director Public Policy Programme, The Alan Turing Institute)

Helen Margetts (Director Public Policy Programme, The Alan Turing Institute, and Turing Fellow, University of Oxford)

Jon Crowcroft (Turing Fellow, University of Cambridge)

Michael Veale (Digital Charter Research Fellow, The Alan Turing Institute)

Thomas Hills (Turing Fellow, University of Warwick)

turing.ac.uk
@turinginst