Legal and public policy considerations: proposal by Global Partnership on AI for collaborative study of TVEC and social media recommender systems

Tom Barraclough, Curtis Barnes

www.brainbox.institute

October 2021

EXECUTIVE SUMMARY

Connection with the Global Partnership on Artificial Intelligence (GPAI)

Brainbox is participating in a project being jointly conducted between the Global Partnership on Artificial Intelligence ("GPAI") and the University of Otago ("Otago"). The purpose of the GPAI project is to lay out a potential methodology for studying the impact of recommender systems used by social media companies. The project also intends to implement this methodology in 2022 if possible by collaborating with a social media company.

Limitations

We emphasise at the outset that this report is intended to canvas the kinds of issues that might arise in connection with the GPAI project. It should not be read as saying that any of these issues are insurmountable, or that they cannot be avoided by careful research design. Further, while we raise matters of public perception, this is not to say that those perceptions are correct, only that they may create disincentives or contextual risks to be aware of when conducting the research and when seeking to establish a collaborative approach.

In this report, we have attempted to constructively contribute to GPAI's proposed approach by sharing insights on the legal and public policy issues raised by the kinds of projects described in the GPAI technical report. We emphasise that the two reports were written concurrently with limited opportunity for interaction between them. It may be that the kinds of issues we raise here can be – or already have been – resolved by the specific research method that GPAI is proposing to adopt.

Nothing in the report should be taken as an argument that social media recommender systems do or do not cause harm: it has been written specifically to avoid adopting a position on this emerging area of research. Instead, our focus has been on considering what kinds of issues might arise if an external party – whether researcher or regulator – were to seek to empirically investigate hypotheses about the impacts of social media companies' systems and their contribution to such harms. We strongly support such empirical investigations. Consistent with our support for those investigations, we have considered as far as we can in the context of this work what might be required to enable them to proceed. Any criticism of GPAI's approach must be assessed in that light.

We also emphasise that, ultimately, the methods employed by the GPAI researchers are technical in nature and better left to those researchers with technical expertise. This report is intended to draw on Brainbox's background in the legal and policy discussions surrounding these technical areas and to summarise these for the benefit of the GPAI researchers.[1]

GPAI project proposes a method for researchers to study recommender systems from outside a platform company

The GPAI project adopts as its starting point the position that it is difficult to study recommender systems in a situation where the researcher is situated outside a social media company.

It is important to briefly note at the outset that there are compelling reasons why external study of recommender systems is so difficult. That is because access to information about recommender systems creates risk from a legal, regulatory, reputational, financial, and commercial perspective. We

[1] This research has been informed by an investigation commissioned by an investor coalition engaging with social media companies in light of the Christchurch terror attacks of 15 March 2019. See: Brainbox (2021). Report to the investor collaboration that has been engaging with social media companies in response to the Christchurch terror attacks. <https://www.nzsuperfund.nz/news-and-media/collaborative-engagement-with-social-media-concludes/>.

expand on this later in the report.[2] We also draw on these insights as part of a wider discussion of moves toward mandated transparency legislation.[3]

It is also important to note that companies' resistance to external scrutiny of their recommender systems need not be explained solely by hostile or antisocial intent. There are a range of structural factors which may make it impossible for platforms to participate in this kind of research, even if they otherwise wished to participate. In particular, there is a complex network of rights and obligations that exists among regulators, users, platforms, contractors, and employees.

Two approaches to studying recommender systems

As a result, if a researcher wishes to study recommender system used by a company, there are two approaches available: voluntary and mandatory. In practice, we believe voluntary approaches will lead to mandatory ones. This is true both in the context of this project, and over the broader global trajectory of platform regulation. In particular, we are already seeing transitions to mandatory approaches.

- The first approach is to take a collaborative approach with a platform company based on mutual agreement with the platforms. Some have tried this, with the most high profile example being the Social Science One Initiative, which faced significant challenges and has arguably failed.[4] There are also a range of multi-party external audit and transparency mechanisms. Because of the increasing calls for transparency, we take it these have failed to satisfy transparency expectations. There are limitations to taking a collaborative approach with a candidate company. Further the results of the work are, in most cases, likely to be compromised (whether as a matter of perception or reality) by a right of veto over publication. As a result, this first collaborative approach will usually inexorably lead to the second approach: compulsory access and transparency through legislation or other legal mandate.
- The second approach is to have a State or group of States mandate a framework for the compulsory disclosure of information sought by outside parties, being information that platforms either do not wish to or cannot disclose. By a framework, we mean a system of rules and principles that reconfigure existing rights and obligations. This will inevitably lead to restrictions or alterations to particular rights or obligations that currently present obstacles to disclosure. These rights and obligations may be held by platforms, their users, their shareholders, and their directors. We say it is preferable to directly acknowledge that these rights and interest are being limited or amended in some way. This allows the various limitations to be imposed following a process of substantive justification. This will involve assembling a body of empirical evidence in order to adhere to human rights principles relating to legitimacy, necessity and proportionality. We expand on this topic later when considering the trade-offs and justifications faced by mandated transparency approaches.

GPAI project proposes a collaborative approach to study recommender systems and TVEC

The Otago GPAI project is aiming to take the first approach: a collaborative and voluntary one. It proposes to test this approach in relation to recommender system behaviour toward a particular class of harmful content: terrorist and violent extremist content ("TVEC"). Among a range of potential classes of harmful content, and particular relationships to be explored, TVEC has been chosen as a

---

[2] See: Second class of issues: issues arising from enhanced transparency, whether collaborative or mandated.
[3] See: Existing legislative proposals for transparency.
[4] 'Public Statement from the Co-Chairs and European Advisory Committee of Social Science One' <https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one> accessed 2 September 2021

particular case study given the opportunity created by commitments to examine algorithmic recommendation systems pursuant to the Christchurch Call and subsequent work plans.

In this report, we set out some of the public policy issues that we anticipate arising from the project GPAI is undertaking. These issues relate to two classes of issues.

First class of issues – adopting TVEC as a subject[5]

The first class of issues stem from adopting TVEC as a subject of study. Studying TVEC produces consequent public policy issues that must be explicitly anticipated by the methodology and the research objectives.

We bring insights to this from recent work we have done touching upon TVEC and content moderation by the platforms.[6] Relevant challenges include definitional issues around what is meant by TVEC and who gets to decide what content is TVEC or not. Studying TVEC also leads to concerns related to the public policy implications of tracking user behaviour in relation to consumption and distribution of alleged TVEC as well as potential discriminatory behaviour.

We think that, by adopting TVEC as a subject of study, the proposal is likely to face many of the issues that arise from proposing content-specific laws or standards for content moderation. In essence, by selecting a class of content (TVEC) the method imports all the definitional issues faced by people designing legal and technical systems for content moderation against TVEC.

There is little we can do to resolve these in this report. All we can do is set them out so that the people looking to pursue the research can factor them into research design and procedure. These issues also illustrate to some extent why the platforms may insist on retaining supervision, oversight, or control over research touching on the platforms' data, technical systems, or operational processes and policies.

Second class of issues – collaborative approach to external study[7]

The second class of issues relate to the collaborative approach being taken and the legal issues presented by taking that approach. Issues of this nature relate to the following.

- Potential risks to user privacy from disclosing data.
- Intellectual property risks from disclosing the details of how algorithmic systems operate.
- Commercial confidentiality considerations from disclosing how business systems operate, including wider operational processes including human decision-making processes.
- Assurance and integrity issues related to potential adversarial or antagonistic engagement with those systems once details are known.
- The risk that selected disclosures about one class of systems leads to unintended disclosures about other classes of systems.
- The rights of third parties who are neither users nor employees, but are contractors or consultants who perform operational functions.[8]

In relation to this second class of issues, we think these illustrate the kinds of factores that are likely to arise when it comes to implementing legislation or regulation that imposes transparency and auditing regimes on social media companies. We lay these out in order to do what we can to contribute to the

---

[5] See below: First class of issues: adopting TVEC as a subject
[6] Above, fn 1.
[7] For a more detailed description, see below: Second class of issues: issues arising from enhanced transparency, whether collaborative or mandated.
[8] Consider for example the rights and interests of external contractors like Accenture or other smaller content moderation sub-contracting firms. See 'How Facebook Relies on Accenture to Scrub Toxic Content - The New York Times' <https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html> accessed 6 September 2021.

efficacy of calls of “transparency” for social media companies, noting these calls have attracted unparalleled global attention with the disclosures of Frances Haugen.[9]

A human rights approach is useful to dealing with both classes of issues

There is a common thread to both these sets of issues. That thread is drawn from human rights law and principles. In previous reports, we have said that regulation of social media platforms is difficult because:[10]

1. We are still beginning to decide, let alone articulate, what the proper role of social media platforms is at a socio-political level globally. This issue creates difficulty when it comes to conceptualising and articulating what the platforms are meant to do and how they are meant to do it in order to meet the expectations of that role.
2. We are still only beginning to articulate, in light of the expected role of social media companies, what the proper role of States is in relation to the platforms, and consequently how the law should be used by states to define the rights and liabilities of the companies, users, and governments as against each other.

By a human rights approach, we mean an approach which starts with a consideration of the relevant human rights which are engaged by a particular research or regulatory proposal. When it comes to imposing regulation which might limit human rights, any limitations should adhere to longstanding principles of legality, necessity and proportionality.

Human rights jurisprudence provides a rich, global and commonly understood starting point for discussion about how to reach conclusions on these issues. This will serve to accelerate the process of identifying points of consensus and dispute between various groups. We apply a human rights approach, so far as we can in the context of this project, to both the first and second classes of issues.

As a result, we hope to illustrate some difficulties of pursuing transparency regulation in relation to social media companies, despite the broad appeal and substantive justifiability of transparency regulation.

[9] ‘Perspective | Facebook Hides Data Showing It Harms Users. Outside Scholars Need Access.’ Washington Post <https://www.washingtonpost.com/outlook/2021/10/05/facebook-research-data-haugen-congress-regulation/> accessed 6 October 2021.
[10] Above, n 1.

WHAT ARE THE KEY FEATURES OF THE GPAI PROJECT?

Before explaining the issues of law and policy we see arising from the GPAI project, we must first outline what that project proposes to do. This report has been written concurrently but separately from the GPAI technical report, although it has been written with the benefit of frequent discussions with the intent that cross referencing can occur between the reports.

Because these reports are intended to be read together to a large extent, here we outline only the core features that we see as being relevant to the legal and public policy issues we propose to discuss. The authoritative statement of the methods and intent of the GPAI project should be taken from the GPAI report.

Key features of the broader GPAI project

The following describe our understanding of the key propositions adopted by the GPAI proposal.

(1) Focus on "recommender systems"

The GPAI project takes as its starting point the role of what are referred to as "recommender systems". It adopts a particular definition for what is meant by a recommender system. It proposes to study how these recommender systems operate within social media companies' systems.

(2) Prima facie cause for concern about the impact of social media recommender systems

The broader socio-political and academic context for the project stems from a concern that the social media platforms use algorithmic systems that have harmful effects. It makes the case that prima facie cause for concern about the operation of these systems has been demonstrated and that such concerns are worthy of further investigation. It is also suggested that such concerns could lead to or justify regulatory intervention, although arguing for regulatory intervention is not the focus of the project itself.

(3) Recommender systems may be harmful because they distribute and expose people to harmful content

The GPAI project rests at least partly on the proposition that recommender systems may be harmful because they recommend content that is harmful. As a result, it proceeds on the basis that harms might be generated by:

- the bare distribution of the content,
- by user exposure to the content, or
- by user responses to being exposed to the content, whether online or offline.

(4) Recommender systems can be harmful because they increase engagement and drive increased consumption of harmful categories of content

Another harm that the GPAI proposal seeks to investigate relates to user responses to recommender systems. The project suggests that harms accrue not just from exposure to the content, but also due to the way that:

- recommender systems increase the consumption of content generally, or
- more specifically increase the consumption of content which is harmful.

It is not just that they expose people to content, but also that people will return to consumer greater amounts of that content, and that harm results from greater consumption of the particular content.

(5) Proposal to assess whether these concerns are empirically justified by studying the recommender systems used by the platforms.

The GPAI project aims to assess whether the prima facie cause for concern discussed in the literature and in public reporting is empirically justified by reference to studying the actual operation of a recommender system inside a social media company.

(6) Proposal to adopt a collaborative approach in order to gain access to key information required for following "internal" methods

The GPAI project outlines that there are methods for assessing recommender systems that are both externally situated and internally situated, from the perspective of where the researcher sits in relation to the company operating the recommender system.

GPAI proposes to adopt a mixture of these approaches: while the research entity will sit outside of the companies, the project anticipates embedding a researcher inside the platform companies. It will do so by adopting a collaborative approach to engaging with social media companies.

The project intends to publish only the results of the experiment, in a way that does not publish user data or confidential commercial information. Specifically, the material for publication will be limited primarily to a scientific paper that shows the strength of the hypothesised relationship.

An important feature of this approach is that GPAI proposes to embed a researcher inside a company who is subject to confidentiality and non-disclosure arrangements. It is anticipated that the company involved will have legal authority to restrict what information can be published by the researchers, whether they are the internally embedded researcher(s) or the external partners to the internal researcher.

(7) Proposal to orient the project toward terrorist and violent extremist content

GPAI proposes to assess the relationships and effects described above in relation to a particular class of content, namely TVEC. TVEC has been selected in part to take advantage of the agreement by tech companies to study algorithmic recommendations and TVEC in the Christchurch Call. This broad agreement in the Call itself has been further adopted within an agreed work plan for 2021 and in public statements:[11]

> *• Working across the Call Community we will develop methods to better understand 'user journeys' and the role algorithms and other processes may play in radicalisation; looking at how the online environment may amplify hatred and glorification of terrorism and violent extremism.*
> *…*
> *• Improving transparency from Governments and Online Service Providers on the policies, actions and tools used to counter terrorism and violent extremist content online;*
>
> *• Continuing to uphold international human rights law and fundamental freedoms online as well as a free, open and secure internet.*

It is this last feature of the proposal – its orientation toward TVEC – that we discuss first in this report in relation the first class of issues.

[11] Public commitments have also been made: see Christchurch Call Second Anniversary Summit – Co-Chair Statement <https://www.christchurchcall.com/second-anniversary-summit-en.pdf> accessed 12 October 2021.

FIRST CLASS OF ISSUES: ADOPTING TVEC AS A SUBJECT

GPAI's proposed method raises a number of challenges from a law and public policy perspective. Some of these challenges may be more a matter of perception rather than reality. Further, they may not be insurmountable, however we believe they do need to be addressed directly.

These challenges are complex. We can only summarise them here while acknowledging the deep knowledge that exists as a result of longstanding work by wider civil society, including NGOs, journalists, and academics.

Item 1: The project must state the specific concerns and the precise relationship to be investigated

When it comes to arguing that the companies should provide external access to their systems, whether to researchers or to regulators, we believe it is essential to state the precise relationship being investigated for both pragmatic and principled reasons. This is because the precise relationship being investigated will have implications for the scope and nature of the access being requested and delivered. The scope and nature of this access will be what dictates the extent and specifics of the legal and public policy issues involved.

There are a range of potential relationships between a range of different systems and actors when it comes to studying the impact of social media products. For example, there is a complex web of potential causative links that could be explored among:

- platform design;
- user upload of content;
- platform moderation of content;
- content distribution, delivery, and consumption;
- user decisions to flag or not to flag content that may infringe content moderation standards;
- the impact on a user as a result of delivering content to them, as distinct from whether they have consumed it in a meaningful way;
- consequent impact on user behaviour, whether online or offline, and whether immediately or cumulatively over a longer time period.

Frequently, public discussion glosses over these nuanced distinctions and overlooks complex distinctions between correlation and causation. In any event, the strength of the evidence and whether it demonstrates correlative or causal links is bound to be a matter of ongoing expert investigation and debate.

There are both practical and principled reasons for specificity in access requests and in the substantive justifications behind transparency and access regimes. We outline some of these below.

Practical reasons for specificity

At a practical level, a researcher or a regulator must be able to frame a request for access to specific information that is sufficiently targeted. Relevant questions include:

- To what exactly is the external body seeking access? We think requests could cover the following loose categories: data (about user behaviour, about system operations, about non-users, or about de-identified users), algorithms and algorithmic systems, user information, internally prepared summaries, operational or procedural documentation, legal opinions and internal advice, contractual arrangements with employees or consultants, or minutes of meetings and discussions.
- Social media companies must be able to comply with the request and this is not possible if the request is overly broad or theoretically endless.

- Practically speaking, if a request is not sufficiently specific, external parties may not receive what they were expecting. That is likely to lead to impressions that companies have not delivered on what was requested, fuelling suspicions of bad faith, regardless of whether or not the companies intended to fulfil the request.
- The relevant legal issues engaged by an access request depend heavily on what information is being sought, who is seeking it, and the purpose for seeking it. These specifics will also trigger different rights for different groups. It is difficult to begin to assess these issues without a specific factual matrix to analyse.
- Connecting the practical and the principled, if we are unable to say what we want the companies to do now through collaborative approaches, it will be difficult to do so later through legislation.

Principled reasons for specificity

The principled reasons for specificity relate primarily to the way that access and transparency regimes are likely to be backed by a legislative mandate in the near future. At a principled level, any regulatory regime will impose access requirements in ways that inevitably impose restrictions on existing rights and interests. To the extent a human rights approach to these restrictions is to be followed, the restrictions should meet the following broad criteria.[12]

- The restriction (the obligation to provide access) should be imposed by law. It should be capable of being reasonably clear and comprehensible. A company should be able to voluntarily comply in good faith. If the legal instrument being imposed on the companies is too vague or discretionary, then it risks becoming an arbitrary enforcement tool that can be used by States for selective and arbitrary enforcement. The principle of legality also requires recourse to legal dispute resolution procedures and rights of appeal in case of disputes about legal interpretation.
- The restriction should be proportional to the harms involved and linked to a legitimate regulatory objective. For example, it is unlikely to be legitimate to impose transparency obligations that could hamper user freedom of expression if the alleged harm relates to identifying anonymous critics who are making true statements about democratically elected public officials. Further, to impose broad and non-specific transparency obligations could lead to onerous transparency obligations out of all proportion to the harms involved with significant implications for citizens.
- In close association with obligations of legitimacy and proportionality, the restrictions imposed on the existing rights and obligations held by companies, users, employees, shareholders and directors must be necessary for achieving the specified regulatory objective. There must be a real link between the restrictions being imposed and the harms being caused, otherwise States could broadly assert there is a link between non-specific conduct and a vague harm and impose restrictions entirely unconnected with the conduct involved. Human rights bodies have concluded there must be a demonstrable connection between the restriction being imposed and the regulatory objective being pursued.

At a principled level, it is not possible to conduct informed democratic debates that weigh and balance competing interests and trade-offs if we have not been adequately specific about what we are proposing to achieve and what we are proposing to alter in order to achieve that. For example, if we fail to recognise that transparency obligations are likely to impact on the rights of contractors or employees, then we are unable to account for these as part of the policy process.

Another point to consider when it comes to weighing and balancing various human rights is that we must be clear about the breadth of the relevant human rights involved. A curious point to consider is the way that international human rights instruments (including the Universal Declaration) do confer

[12] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35, Human Rights Council, Thirty-eighth session (18 June – 16 July 2018).

some significance on rights to property, the right to be protected by the law from attacks on honour and reputation, and rights to intellectual property. This is not to say those rights are absolute or to express a personal view on how they should be weighed against matters like personal safety, to privacy, to freedom of conscience, and freedom from incitement to violence and discrimination. Our point is that all of these rights are engaged at some level by proposals to impose transparency requirements and the current global discussion about social media products. We cannot credibly weigh and balance them if we fail to realise they are engaged.

Item 2: Using natural language rules and principles to classify/categorise content is conceptually and linguistically challenging

It is difficult to use natural languages (English, for example) to delineate between categories of content or human expression which should be permitted or prohibited. One issue facing the GPAI project is that, even before we reach the point of computationally classifying content, it is difficult to say what is TVEC and what is not. Further, the best-efforts attempts adopted by governments and by civil society bodies involve the use of language that necessarily requires assessments of value, judgement, and degree. This will not be true in all cases: for example, it is difficult to reasonably argue that content like the Christchurch livestream videos is not TVEC. Some cases will be clear-cut. But the point is that generally speaking, it is difficult for the researchers and others to say what they mean when they refer to the category "TVEC".

The activity of placing content into categories is often referred to as "classification". "Classification" is described by Gorwa et al[13] as follows:

> *Classification … assesses newly uploaded content that has no corresponding previous version in a database; rather, the aim is to put new content into one of a number of categories.*

"Classification" is an ambiguous word in a legal/technical context because it has different legal and technical meanings.

- Computational classification refers to a task performed by computational systems (usually associated with machine learning). In a content moderation context, particularly in relation to TVEC, machine learning is used to make (often) probabilistic assessments about whether content breaches content moderation standards.[14] Specifically in relation to TVEC, the companies sometimes use face and voice recognition to identify known terrorists, as well as symbol identification.
- Classification is also a sociolegal activity conducted by censorship or content moderation bodies to make judgements about whether content sits in a particular category or not. The language of "classification" is still used in relation to age restrictions on movies, for example, and is reflected in the title of New Zealand's censorship legislation.[15]

The key point of item 2 is to recognise that even without the added complexity generated by the use of computational systems, content classification is a difficult exercise. This makes it difficult for the

---

[13] Gorwa R, Binns R and Katzenbach C, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945

[14] A defining event in the history of this practice is the genesis of the Syrian Archive organization, which resulted from increased use of ML classification by YouTube to automatically remove TVEC, and the destruction of evidence of war crimes in Syria. The use of such automated systems also frequently has discriminatory effects for all the usual reasons.

[15] The Films, Videos and Publications Classification Act 1993.

researchers to say what content they are targeting when they refer to TVEC. This compounds the difficulties referred to in our discussion of item 1.[16]

We offer two examples illustrating how the articulation of natural language rules that are capable of sorting human expression into different categories is difficult. The first comes from a private ruleset, Twitter's community guidelines. The second comes from legislation that pre-dates the ascendancy of the social media companies, New Zealand's Films, Videos, Publications and Classification Act 1993.

Example: Twitter rules on violence

TVEC (terrorist and violent extremist content) is actually covered by a range of different community standards, not just one. One of the relevant standards on Twitter relates to extreme violence.

Terrorism-based standards have historically relied on a State-designation process before a group can be identified as "terrorist". This means lone actor attackers or de-centralised extremist movements may not meet the requirements for terrorism-related standards. Relying on violence-based standards to target terrorist content is one way to avoid such issues in a way that would, independently, justify the removal of the content. Twitter's content moderation standards state:

> *Violence: "You may not threaten violence against an individual or a group of people. We also prohibit the glorification of violence."*

Against this broad standard, there are immediate exceptions, including:

> *We recognize that some people use violent language as part of hyperbolic speech or between friends, so we also allow some forms of violent speech where it's clear that there is no abusive or violent intent, e.g., "I'll kill you for sending me that plot spoiler!". … Under this policy, you can't glorify, celebrate, praise or condone violent crimes, violent events where people were targeted because of their membership in a protected group, or the perpetrators of such acts. We define glorification to include praising, celebrating, or condoning statements, such as "I'm glad this happened", "This person is my hero", "I wish more people did things like this", or "I hope this inspires others to act".*

The most notable exception to this, which explains why TVEC can be a politically fraught area, is for cases where a user praises extreme violence conducted by a State:

> *"Our focus is on preventing the glorification of violence that could inspire others to replicate violent acts, as well as violent events where protected groups were the primary targets or victims. Exceptions may be made for violent acts by state actors, where violence was not primarily targeting protected groups."*

Example: New Zealand Films Videos Publications Classification Act 1993

Another illustrative example is drawn from State-based legislation. New Zealand's Films Videos Publications Classification Act 1993 confers powers on the Office of Film and Literature Classification to "classify" publications, which include a broad array of audio-visual or textual objects.

While localised to New Zealand, this law does have some global impact given it makes possession and distribution of the Christchurch livestream and related publications a criminal offence punishable by imprisonment. Publications can be classified as objectionable pursuant to section 3 and dealing in objectionable publications is a criminal offence that can lead to imprisonment. A publication is given an "objectionable" classification following a decision-making process. It broadly relates to publications depicting child sexual exploitation material, terrorist or extremely violent content, and matters such as bestiality.

The classification process to be followed is a staged one that directs attention to both what a publication depicts or describes, as well as whether it "promotes or supports, or tends to promote or support" particular harmful outcomes. In making that assessment, various matters "shall also be

---

[16] Item 1: The project must state the specific concerns and the precise relationship to be investigated

considered" which go to the likely context and impact of a publication in context. The articulation of these various contextual matters in a statute illustrates how the actual semantic significance of content cannot be realistically assessed in isolation from its context, including:[17]

> *(a) the dominant effect of the publication as a whole:*
>
> *(b) the impact of the medium in which the publication is presented:*
>
> *(c) the character of the publication, including any merit, value, or importance that the publication has in relation to literary, artistic, social, cultural, educational, scientific, or other matters:*
>
> *(d) the persons, classes of persons, or age groups of the persons to whom the publication is intended or is likely to be made available:*
>
> *(e) the purpose for which the publication is intended to be used:*
>
> *(f) any other relevant circumstances relating to the intended or likely use of the publication.*

In summary, it is difficult to use language to create a set of categories that delineate between permissible and impermissible content. These difficulties are compounded by the semantic complexity of the content itself, which may have radically different "take-out" meanings depending on the context in which it is presented, including temporal context (the same thing said at a different time may have different significance) and based on the known features of the speaker.[18] Even where these rules and principles are reduced to natural language, they will inevitably require a degree of interpretation and judgment in the way they are applied in specific cases. All of these things amplify the scope for reasonable argument about whether a given piece of content does or does not fall into a particular category.

The effect of this for the GPAI project is that it complicates the exercise of articulating clearly what kind of content GPAI proposes to study,[19] even before any questions of algorithmic classification and detection arise.[20]

Item 3: The research project is proposing a category we understand as "TVEC-adjacent" content.

The research project proposes to articulate a new category of content which is simultaneously not-TVEC, but remains sufficiently connected to TVEC such that it is worthy of identification, tracking and scrutiny. This effectively proposes a new category of content in addition to what is already a fraught definitional area, as described in item 2.[21] It is not clear to us how "TVEC-adjacent" content can be adequately defined. We note that, because TVEC is already banned from the platforms, the project's real focus is on this TVEC-adjacent category, not TVEC itself. Therefore the ability to define this category is crucial.

Examining content that is "TVEC-adjacent" (but not TVEC) raises issues from a legal and public policy perspective.

1. To the extent that the project proposes to identify such content by tracking user journeys, this will raise concerns about privacy and about potential discrimination. For example, how strong does the link need to be between "not-TVEC" and the likelihood of subsequent consumption

---

[17] Films, Videos and Publications Act 1993, s 3.
[18] Slurs and terms of abuse can be empowering or objectionable depending on who is speaking and to whom, with some terms being reclaimed by black Americans and by people who are LGBTQI.
[19] Related to Item 1: The project must state the specific concerns and the precise relationship to be investigated.
[20] Item 7: Challenges arising from automated classification of content.
[21] Item 2: Using natural language rules and principles to classify/categorise content is conceptually and linguistically challenging

of TVEC before it is unjustifiably discriminatory to link such content with TVEC? One risk that civil society groups will anticipate is that, for example, content in Arabic or that discusses Islam might be more likely to be classified as TVEC-adjacent than other kinds of equally qualifying content.

2. There is a real risk that a conclusion that "people who eventually watch TVEC also watch this" could lead to a conclusion that "people who watch this are likely to commit a TVEC-related crime". Once the platforms become aware of this knowledge, what should they do with it? This raises ethical risks. It is also worth considering, in light of recent events, how such work could be reported on publicly by news media and how regulators and the public might react if they knew such knowledge existed, or such research was taking place.
3. This raises questions of temporality and proximity. It also encourages researchers to draw conclusions from an incomplete dataset – i.e. the user's total information exposure cannot be known.[22] There is no means for researchers to know what information a user consumed elsewhere on the internet, or in physical books, or on television, or in conversation. Any of this information might have had a causative effect on the user consuming TVEC at some later point, potentially greater than the information consumed by the user on the platform.
4. From a law and public policy perspective, generating research that could reasonably support a conclusion that someone could commit a TVEC-related offence could be difficult for the platforms who, in some jurisdictions, have obligations to report TVEC-related conduct to law enforcement bodies.
5. Another point that must be considered is the way that the platform companies have collaborative relationships with civil society groups built on the premise that transparency and trust are important: specifically, the GIFCT. A core concern held by civil society groups is that the platforms may become unjustifiable and unaccountable surveillance and discrimination tools for nation states seeking to persecute legitimate political opponents or minority groups. We anticipate that a category of "TVEC-adjacent" content, if not adequately defined, could cause concern to these groups.

Civil society groups collaborating with groups such as GIFCT are concerned about the risk that gradually broadening definitions of TVEC can lead to greater intrusions in areas that were previously off-limits to State restriction. The GIFCT transparency working group recently published a report expressing concerns about "scope creep":[23]

> *"Definitions and Concerns of Scope Creep: Participants pointed out the need to carefully parse out what scope creep is. GIFCT work is not about "objectionable" or wider "harmful" content. It is about definable terrorist and violent extremist content. There are cultural sensitivities and no agreed definitions for "extremism" or "radicalized" content."*

It is very likely that proposing a TVEC-adjacent category will trigger similar concerns.

Item 4: The proposal will need to adopt a definition of TVEC from a range of possible sources, which could generate objections

Taking TVEC as an example, there are a range of different institutional definitions for articulating the natural language rules and principles referred to in item 2. These definitions could be drawn from:

---

[22] This is one issue faced by the Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019 given the attacker's use of virtual private networks and other technologies for obscuring his internet behaviour. They faced this difficulty even with some access to the resources and insights from investigations by national security and intelligence agencies.

[23] 'GIFCT Transparency Working Group: One-Year Review of Discussions' (2021) <https://gifct.org/wp-content/uploads/2021/07/GIFCT-WorkingGroup21-OneYearReview.pdf> accessed 7 September 2021 at p 11.

- Companies' respective community moderation standards (bearing in mind these are frequently supplemented by detailed decision-making guidelines and decision-tree documents for the actual moderators applying them).
- Definitions by multilateral institutions that include States, such as Tech Against Terrorism or the EU Codes of Practice on disinformation.[24]
- Definitions by multilateral institutions that do not include States, such as GIFCT or the Global Network Initiative.
- Civil society groups.
- Academic writing.
- State-level legislation and regulation, for example from the European Union or from individual States. Specific examples might include the Australian Abhorrent Violent Material amendments, or the EU 24-hour terrorist content law.

The research project may have to adopt one of these definitions and this will carry with it political and legal implications. What we mean is that the researchers will be perceived to have adopted a position on what should or should not be categorised as TVEC by choosing one definition over another, and thereby to have committed to a political position. This might raise issues of perception both for the research project itself and for the candidate company participating in the collaboration.

It seems likely to us that the project will have to adopt a definition based on the definition which is operationalised by the platform companies. The companies' own definitions are likely to be embedded in their existing systems and reflected in their existing datasets. This is likely to import all of the strengths, biases or limitations incorporated into a specific company's given approach and thereby expose the research project to those same criticisms.

Item 5: Classifying content using natural language rules and principles raises significant public policy issues around human rights like freedom of expression.

Decisions to articulate categories of human expression and to apply them occur against the background of the human right to freedom of expression. The human right to freedom of expression is not absolute and rights to expression can be limited by State-level legislation for a range of justifiable reasons, including to protect the rights of others.

The GPAI project will be required to adopt a definition of TVEC (and of TVEC-adjacent content) which will engage legal and policy issues around freedom of expression. These can only be assessed once a particular categorisation has been adopted, in light of the issues we raise at items 2-4. The GPAI project should be aware of this as it works toward taking a collaborative approach.

Item 6: Among a range of possible subjects, TVEC is one of the most politically and conceptually contested

Political and legal implications of categorising TVEC

To people not acquainted with the literature around politics, international relations, law and other related areas, the definition of "terrorist and violent extremist content" can seem clear cut and easy to

[24] 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (European Commission - European Commission) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 1 April 2021.

apply. By contrast, it is one of the more hotly contested areas that could be used to explore the issues related to content moderation and freedom of expression.[25]

The consequences of a State labelling an individual or a group as a terrorist or violent extremist organisation are significant. When a State can successfully allege that a person or group fall into the category of terrorists or violent extremists, they can leverage the full power of the State to interfere in that group's operations and expression. Further, if a State identifies any legal limitations on its ability to respond to a terrorist threat, then it can use the risk of violence to justify broad erosions of civil liberties and human rights to privacy, due process and freedom of association. In fact, this reactive limitation on civil liberties by States in the name of safety is one of the key harms that can flow from a terrorist incident.

States also use violence and suppress human rights. Because States also use violence for political purposes, the task of framing what is and is not "terrorist" or "violent extremist" behaviour inevitably imports assumptions about when the use of violence or the use of expression that legitimises violence is acceptable.

This makes the definition a qualitative and value-laden one to apply, even if the narrowest possible definitions are adopted. For example, even the commission of violent acts by authorised State actors acting within domestic law can be a grave abuse of human rights that might ordinarily meet a plain language definition of terrorism.

GIFCT taxonomy for shared hash database and common definitions

The relevant issues we have grouped under items 2,[26] 3,[27] 4,[28] 5[29] and 6[30] are exemplified by the GIFCT's attempt to find a common definition of "TVEC".

The GIFCT recently commissioned a human rights impact assessment report on itself and its key operational features as an organisation. The human rights impact assessment dealt with a question related to whether definitions of TVEC could be harmonised across GIFCT or across the platforms. The report concluded:[31]

> *We note the debate about competing and conflicting definitions of terrorist content, with concern that some existing government definitions of terrorist content may encompass legitimate expression protected under international human rights law. Overall, this assessment surfaced widely-held skepticism toward GIFCT creating a shared definition of terrorist and violent extremist content, and BSR has come to agree with this sentiment. We believe this task properly resides with governments, and acknowledge both the huge challenge of reaching consensus and the fact that companies operating in very different contexts may need different definitions.*
>
> *However, while stopping short of a shared definition, we do believe that creating a common understanding of terrorist and violent extremist content—even if companies choose to adapt their own precise definitions—would have considerable value. Benefits would include (1) pushing back against overbroad definitions of terrorist and violent extremist content deployed by governments; (2) improving the capability of smaller companies without extensive policy teams to*

---

[25] We note the summary of various kinds of harmful content in part 2 of the GPAI technical report.
[26] Item 2: Using natural language rules and principles to classify/categorise content is conceptually and linguistically challenging
[27] Item 3: The research project is proposing a category we understand as "TVEC-adjacent" content.
[28] Item 4: The proposal will need to adopt a definition of TVEC from a range of possible sources, which could generate objections
[29] Item 5: Classifying content using natural language rules and principles raises significant public policy issues around human rights like freedom of expression.
[30] Item 6: Among a range of possible subjects, TVEC is one of the most politically and conceptually contested
[31] 'Human Rights Impact Assessment: Global Internet Forum to Counter Terrorism | Reports | BSR' <https://www.bsr.org/en/our-insights/report-view/human-rights-impact-assessment-global-internet-forum-to-counter-terrorism> accessed 6 September 2021.

> *establish their own definitions; (3) establishing a bulwark against "slippery slope" definitions of terrorist and violent extremist content that may extend too far into other forms of speech, thereby presenting risks for freedom of expression; and (4) improving shared awareness of the relationship between human rights and terrorist and violent extremist content. In addition, we note that the multi-stakeholder setting of GIFCT is an excellent opportunity to improve understanding, generate increased consensus, and enhance the shared mission of GIFCT participants. BSR recommends starting with a common understanding of terrorist content, then moving on to violent extremist content, which presents a more complex challenge given its adjacency with broader notions of hate speech and extremist (but not violent) content*.

By way of illustration, the BSR point to a range of "models of a shared definition", including:

> *Article 3(1) of the EU regulation on preventing the dissemination of terrorist content online or the definition proposed by the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism as existing definitions to adopt—while others had in mind the much longer and detailed definitions that are actionable by companies.*

To illustrate the kinds of features involved in defining TVEC, we reproduce BSR's summary of art 3(1) of the EU regulation:

> *… terrorist content means one or more of the following: (a) inciting or advocating, including by glorifying, the commission of terrorist offenses, thereby causing a danger that such acts be committed; (b) encouraging the contribution to terrorist offenses; (c) promoting the activities of a terrorist group, in particular by encouraging the participation in or support of a terrorist group; (d) instructing on methods or techniques for the purpose of committing terrorist offenses.*

BSR recommended that a multi-stakeholder approach be adopted for development of relevant definitions and taxonomies. Its proposed stakeholder exercise illustrates just how wide-ranging and complex such an exercise can be:

> *BSR proposes that the common understanding be created in a collaborative and multi-stakeholder manner, including the new IAC taxonomy subgroup, relevant working groups, and Global Network on Extremism and Technology (GNET), the academic research arm of GIFCT GNET research. There should be consultation with relevant UN Special Rapporteurs, dialogue with affected stakeholders, and transparency about how final resolutions are reached. These common understandings should be written to a level of granularity that is actionable and practical for companies to use, including, but not limited to, requirements for content added to the hash-sharing database. While each company will likely retain its own definition, these common understandings can shape and inform the work of GIFCT. They should be reviewed on a regular basis to ensure they accurately reflect reality and adequately limit overreach.*

Categorisation of TVEC is usually contestable and fact specific

Even if a definition can be agreed, the question of whether a given piece of content meets that definition is a complex one. A feature of assessing whether content is or is not TVEC is that it frequently relies on the ability to assess the intent or objective of the speaker. This is difficult because it requires nuanced evidential assessments of the statements made (or implicit in the content), as well arguments about the likely conclusions that will be drawn subjectively in the mind of an audience.

Because the categorisation of content as TVEC or not-TVEC is usually arguable, this creates an opportunity to skew qualitative assessments of content depending on whether or not the categoriser wishes to restrict the content or not. In essence, someone wishing to silence someone can take a broader or narrower interpretation of their words, where a more charitable interpretation might have prevented that restriction.[32] Depending on the approach of the enforcement body, the qualitative

[32] Susan Benesch has suggested an empirical methodology for the platforms to assess how a statement that may be an incitement to violence is being understood by people based on assessing user engagement in response to a post: see the dangerous speech project led by Susan Benesch. 'Op Ed: To Keep Social Media from Inciting Violence, Focus on Responses to Posts More than the Posts Themselves | Dangerous Speech Project' (Dangerous Speech Project |, 31 May 2021)

breadth of incitements to violence can work either in favour of the inciter or in favour of the moderator. People calling for violence or genocidal actions may seldom directly call for them in unambiguous terms, meaning their words must be construed in context and a complex argument made about why they are an incitement.

When it comes to platform assessments of incitement to violence, the context may include statements made outside a platform. This approach informed the Oversight Board's analysis of Facebook's decision to suspend President Trump:[33]

> *To understand the risk posed by the 6 January posts, the Board assessed Mr Trump's Facebook and Instagram posts and off-platform comments since the November election. In maintaining an unfounded narrative of electoral fraud and persistent calls to action, Mr Trump created an environment where a serious risk of violence was possible. On 6 January, Mr Trump's words of support to those involved in the riot legitimised their violent actions. Although the messages included a seemingly perfunctory call for people to act peacefully, this was insufficient to defuse the tensions and remove the risk of harm that his supporting statements contributed to. It was appropriate for Facebook to interpret Mr Trump's posts on 6 January in the context of escalating tensions in the United States and Mr Trump's statements in other media and at public events.*
>
> *As part of its analysis, the Board drew upon the six factors from the Rabat Plan of Action to assess the capacity of speech to create a serious risk of inciting discrimination, violence or other lawless action[, including]: context …; status of the speaker …; intent …; content and form …; extent and reach …; imminence of harm ….*

The Trump decision also illustrates why the definition of TVEC and its application in any specific case is so hotly contested: the authority to label President Trump's actions or his supporters' actions as being TVEC-related could lead to a range of outcomes that authorise state use of power against an elected president.

Contemporary examples

A contemporary example (among many) is the incident in May 2021 when a plane travelling from Greece to Lithuania was grounded in order to detain Roman Protasevich.[34]

> *Mr Protasevich is a former editor of Nexta, a dissident media operation with a popular Telegram messenger channel. He left Belarus in 2019 and has been living in exile in Lithuania. Nexta became a significant channel for protesters challenging the August 2020 presidential election in Belarus, widely condemned as rigged.*

Protasevich has been dealt with using laws directed toward terrorism and inciting unrest, according to the BBC:[35]

> *Mr Protasevich faces charges of organising mass unrest after covering the events of the 2020 presidential election from abroad. The offence carries a possible jail term of up to 15 years. However, Mr Protasevich tweeted a KGB list of terrorism suspects last year, adding that he had been placed on it alongside Islamic State jihadists. Terror offences reportedly carry the death penalty in Belarus. Opposition figures and independent journalists have been arrested in large numbers in Belarus in recent months. Human rights group Viasna reports that 472 people are being held as political prisoners.*

Another recent example is the way that an "extremist" designation was applied to Alexei Navalny and his "Anticorruption Foundation" in Russia. The organisation had created a "smart voting" app to

---

<https://dangerousspeech.org/to-keep-social-media-from-inciting-violence-focus-on-responses-to-posts-more-than-the-posts-themselves/> accessed 7 September 2021

[33] 'Oversight Board, Case Decision 2021-001-FB-FBR' (6 May 2021): <https://www.oversightboard.com/decision/FB-691QAMHJ> accessed 7 September 2021

[34] 'Belarus Plane: What We Know and What We Don't' BBC News (25 June 2021) <https://www.bbc.com/news/world-europe-57239521> accessed 7 September 2021

[35] Ibid.

enable strategic voting in federal elections. The extremist designation led to the “smart voting” app being removed from Google and Apple’s app stores. It also allowed authorities to threaten prosecution against Apple and Google and its employees personally for refusing to remove it. Extremist designations also expose users to risk of prosecution if they disseminate information connected with that organisation.[36]

The question of whether TVEC can be defined and how that definition should be framed is a live issue for the tech companies and for collaborative bodies such as GIFCT.

Item 7: Challenges arising from automated classification of content

The GPAI report notes that its proposed method relies on the use of automated classifiers to categorise content. It also notes the limitations of these methods. Here, we give a generic overview of the kinds of legal and public policy issues that are likely to be engaged by the use of automated classifiers in the context of TVEC.

Automated classification may be inaccurate or discriminatory

The core difficulty with any algorithmic classification of content is that algorithmic systems find it difficult to assess the complete semantic meaning of a statement as it might be understood by an informed person in a particular social context. Generally, algorithmic systems are restricted to assessing the content itself. In this sense, they perform poorly when it comes to assessing how context creates and alters the meaning of a statement as it might be understood by a user. This task is complicated further when the content being classified is audio-visual material like pictures or videos.[37] Classifying content as falling within the TVEC category can rest heavily on the ability to assess these contextual indicators. As discussed, classification decisions can also rely heavily on inferences about the likely intent of the content creator based on a range of features of the content itself and its context.

The matter of algorithmic false positives and false negatives will be well known to GPAI, but the diagram below may be helpful to policymakers. These are usefully outlined below:[38]

	CLASSIFIED AS NOT HARMFUL	CLASSIFIED AS HARMFUL
CONTENT WHICH IS HARMFUL	**False negative** **Incorrect classification** Harmful content is not removed, leading to harm to viewers and damage to platform’s reputation	**True positive** **Correct classification** Content correctly removed
CONTENT WHICH IS NOT HARMFUL	**True negative** **Correct classification** Content correctly remains online	**False positive** **Incorrect classification** An ineffective application of the platform’s T&Cs in which content is removed when it shouldn’t have been, possibly curtailing freedom of expression and damage to platform’s reputation

Figure 18 – Content moderation errors can be made in two ways and these have different consequences (**SOURCE:** Cambridge Consultants)

[36] ‘The Lawfare Podcast: Russia Cracks Down on Social Media’ (Lawfare, 7 October 2021) <https://www.lawfareblog.com/lawfare-podcast-russia-cracks-down-social-media> accessed 10 October 2021

[37] Explored in detail in Shenkman C, Thakur D and Llansó E, ‘Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis’ (Centre for Democracy and Technology 2021).

[38] Cambridge Consultants, on behalf of Ofcom UK, ‘Use of AI in Online Content Moderation’ (2019) <https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf> accessed 7 September 2021

The European Commission has adopted a proposal to impose legal requirements on platforms to remove TVEC within 24 hours. This proposal has caused significant concern among human rights groups. Some of these concerns rest on the short compliance timeframes available to the platforms and the way that this will induce reliance on automated systems for detecting and removing such content. A joint letter was produced and sent by three UN special rapporteurs: on freedom of expression, privacy and counter-terrorism. That letter gives a concise overview of the likely legal and public policy issues raised by undue reliance on automated classification systems.[39]

> *The Special Rapporteurs also note that the use of automated tools for content regulation, as required under the draft Regulation, comes with serious limitations and aggravates the risk of pre-publication censorship. Algorithms frequently have an inadequate understanding of context and many available tools, such as natural language processing algorithms, do not have the same reliability rate across different contexts. They have, at times, also been shown susceptible to amplifying existing biases. Moreover, considering the volume of user content that many hosting service providers are confronted with, even the use of algorithms with a very high accuracy rate potentially results in hundreds of thousands of wrong decisions leading to screening that is over- or under-inclusive. The Special Rapporteurs note that Article 9 of the Proposal requests Internet platforms making use of automated tools to provide "effective and appropriate" safeguards to ensure that decisions taken pursuant to the Regulation are "accurate and well-founded". They however wish to highlight that ensuring accurate and well-founded decision-making involving the use of automated tools requires a human rights-based approach to be at the centre of the design, deployment, and implementation of artificial intelligence systems. They further contend that such systems must be subject to human rights impact assessments, periodic independent audits, safeguards ensuring adequate user notice and consent, and robust oversight, including human oversight, of their functioning and use.*

In summary, the accuracy of automated content classifiers is already of concern across a range of human rights-related areas and these concerns will be brought to bear on the GPAI project.

The accuracy of automated content classification tools has acquired a heightened public significance in light of the allegations made by Frances Haugen about the accuracy of Facebook's AI systems for detecting violating content. One of Haugen's core concerns relates to the way that recommender systems can and will recommend content that violates content moderation standards, but has been missed by automated classification tools:[40]

> *Facebook says "We can do it safely because we have AI. The Artificial intelligence will find the bad content that we know our engagement-based ranking is promoting". They've written blog posts on how they know engagement-based ranking is dangerous, but the AI will save us. Facebook's own research says they cannot adequately identify dangerous content and as a result those dangerous algorithms that they admit are picking up the extreme sentiments, the division, they can't protect us from the harms that they know exist in their own system.*

This heightened awareness of the accuracy of content moderation classifiers – such as those for detecting TVEC – will be one matter relevant to public perception of the GPAI project and the companies' willingness to pursue a collaborative approach.

---

[39] David Kaye, Joseph Cannataci, Fionnuala Ni Aolain, 'Mandates of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression; the Special Rapporteur on the Right to Privacy and the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism. Letter Regarding the European Commission's Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online to Complement Directive 2017/541 on Combating Terrorism. OL OTH 71/2018' (7 December 2018) <https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234> accessed 6 September 2021

[40] C-SPAN, Facebook Whistleblower Frances Haugen Testifies before Senate Commerce Committee (2021) <https://www.youtube.com/watch?v=GOnpVQnv5Cw> accessed 12 October 2021

Automated classification may not reflect natural language categories adopted

In items 2-6, we have outlined a range of issues that mean it is difficult to articulate clear categories for content in natural language. Even if these conceptual issues can be resolved, another issue is that the project rests on the ability to use computational systems to identify content in ways that reflect these categories. There is a risk that the automated classification system being used does not actually reflect the natural language categories adopted by the research team. This may undermine the ability to make claims about user engagement with TVEC or user journeys toward TVEC, because the material being identified by the system may be different from the content that categorisation is intended to identify.

We also urge caution about any suggestion that the GIFCT shared hash database can be used as a kind of proxy definition of TVEC shaped by practice rather than *ex ante* rule-making. We urge caution about this for the following reasons:

1. It is likely this would be an expanded functional use of the shared hash database. This is unlikely to be accepted by civil society groups and may jeopardise companies' relationships with those groups. The gradually expanding use of the shared hash database is already a matter of concern to academic commentators and civil society groups.
2. The shared hash database predominantly includes audio-visual material and may not include text-based material. It was recently announced that terrorist manifestos would be added to the database but this may not yet have been implemented. Using the database as a kind of definition will therefore exclude a whole class of content which is likely to be of interest to the researchers.
3. There is a reasonable prospect that there is no central register of what content is held in the shared hash database. We understand that the processes for adding content to the database are ad hoc and unique to each company. One issue faced by the companies is that there can be legal consequences for them if they store copies of the illegal content they are adding to the database. It may be that actual records of what a hash refers to have been deleted once the hash has been added to the database.
4. The shared hash database stores hashes, but not the content itself. The content cannot be recreated from these hashes. Any use of the shared hash database would therefore have to perform some kind of comparison between the hash of content on the platform and a hash in the shared hash database. This will need to be incorporated into research design.
5. It is increasingly clear that the hashes added to the database have been produced through perceptual hashing techniques, which aim to identify matching content according to relevant semantic features, rather than the fact they are digitally identical. This may raise issues of false positive and false negative detections.
6. There is no identifiable or externally scrutable quality assurance process for verifying the legitimacy of a decision by any one company to add content to the shared hash database. This has undermined public perception of the legitimacy of the shared hash database. This ought to be considered if GPAI is proposing to treat the database as a good indicator of definitional consensus or common practice.
7. Some of the hashes added to the database have been added under crisis conditions, particularly during the activation of the GIFCT Content Incident Protocol, used to respond to live terrorist incidents in Christchurch, Halle and Arizona. There is therefore a risk that hashes have been added erroneously, or by way of over-inclusion, in order to err on the side of user safety.
8. The composition of the shared hash database is weighted toward particular CIP incidents where high volumes of content was produced as part of an attack. Specifically, there is a much higher quantity of content connected to, for example, the Christchurch attack (6.8%) rather than the Halle (2%) and Arizona (0.1%) attacks.[41]

[41] See GIFCT Transparency Report for 2020: <https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf>.

9. Anything in the shared hash database is likely to have been removed from the platforms already using automated matching techniques. This may mean hashed content has limited value for mapping user journeys.

To summarise, the use of the GIFCT shared hash database as a kind of "definition by practice" is the exact kind of scope creep that civil society groups are concerned about in relation to both the use of the shared hash database, the role of GIFCT, and the expanding definitions of TVEC. Further, there are a range of factors to consider that may undermine any suggestion that the shared hash database is either substantively or procedurally legitimate as an indicator of an emerging consensus definition.

Statistical proxies for the law may be illegitimate

There is another emerging view which we believe is important for GPAI to consider. We associate this view with the work of the CoHUBICOL project, funded by the European Research Council and led by Professor Mireille Hildebrandt.[42]

We understand Hildebrandt and her collaborators to argue that there is a fundamental difference between articulating what the law is through statistical methods, as opposed to the principled doctrinal reasoning methods followed by lawyers. Hildebrandt asserts that the doctrinal reasoning methods used to arrive at a correct interpretation of the law are closely linked to law's history as a tool for imposing limits on executive power through the technology of written text. Using statistical comparison in order to arrive at a correct understanding of the law fails to apply the method that gives legal interpretation its legitimacy, and therefore statistically derived interpretations lack that legitimacy. We think these concerns are engaged by any suggestion that:

- a legal definition of TVEC can be constructed by statistical comparison with the existing shared hash database; and
- that automated classifiers which use statistical methods are any reflection of the natural language categories adopted by the researchers following items 2-6.

We think that Hildebrandt's concerns are not alleviated by the literature that questions whether human decision-making is really as scrutable as we think it is when it comes to automated decisions.[43] Hildebrandt's concerns are wider than whether a decision-maker is giving adequate reasons for the decision they have made. They relate to whether the interpretation of the law embedded in that system has any legitimacy. The outcome is not that the system is wrong in fact, it is that the system is wrong in law, because it has not applied a correct understanding of what the law requires. It is about legitimising an interpretation of the law that has been adopted by reference to reasons that the law recognises as legitimate. This is an emerging and complex area of jurisprudence which we refer to here and follow with interest.

Reputational and commercial risk from scrutiny of related systems

One point acknowledged in the GPAI report is that the accuracy of the existing content classification systems used by the companies may be very low. Further, these issues with the accuracy of classification systems may be present not just in content ranking algorithms, but may also apply to assessments of user interests and ad-targeting. We think this is an important point to consider from a legal and public policy perspective.

It remains possible that, in reality or as a matter of perception, the companies' core systems for delivering targeted content to users based on their interests are actually not that good. If that was the case, they would be extremely cautious about disclosing insights into how those systems work

---

[42] See a recent COHUBICOL working paper: <https://publications.cohubicol.com/working-papers/text-driven-normativity/>. See also Hildebrandt M, 'The Adaptive Nature of Text-Driven Law' (November 2020) Journal of Cross-disciplinary Research in Computational Law.

[43] Zerilli J and others, 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?' <https://www.repository.cam.ac.uk/handle/1810/299973> accessed 29 September 2021.

because it would undermine the core product they're selling.[44] It was frequently noted in response to the Cambridge Analytica incident that the actual contribution of microtargeting ads towards behaviour (let alone how they vote) cannot actually be shown. In Frances Haugen's testimony before the United States Congress, the question of fraudulent claims about advertising was also raised by a member of the committee.

Item 8: The content being investigated is already contrary to platform terms of service. Distributing that content may also be illegal.

In practice, there are very narrow categories of content where bare possession of it or distribution of it creates a harm significant enough that it will attract legal consequences. These categories include child sexual abuse material (CSAM), terrorist propaganda or extremely violent material, and the distribution of material which is subject to intellectual property rights, or is private or confidential. In practice, the platforms already have a range of systems for restricting the distribution of such content. These systems presumably work by removing that content from the pool of content available to a recommender system for recommendation to a user.

This relationship has been a key point emphasised by Frances Haugen. Haugen has pointed out more than once her opinion that the information she has released, or that Facebook otherwise holds, shows:

- Integrity systems, or content moderation systems, are largely automated, and this automation is inadequate for catching all infringing content. Further, some systems are simply unable to operate on content in particular human languages, or where human content moderators are not available.
- Recommender systems will recommend content that is likely to breach social media company guidelines. Because of technical and administrative deficiencies in integrity and content moderation systems, recommender systems have a tendency to recommend violative-but-missed content to users.

Another key point to consider is that it creates legal and reputational risks for the platforms to accept that such material is present on their platforms and they certainly cannot leave it on the platforms if it has been algorithmically identified: in some jurisdictions this would lead to legal penalties; in others, it would drastically increase political interest in regulating content moderation.

We note that the GPAI technical report outlines suggested methodologies that nevertheless can be used to assess the core dynamic they wish to investigate – a user journey from not-TVEC, to TVEC-adjacent, to TVEC. Nonetheless, the methodology must be designed to anticipate the contextual factors we raise here. It may be difficult to communicate any limitations on the findings that result and this will be another point of risk anticipated by the companies.

Item 9: The project proceeds on the basis that user exposure to legal, non-violative content might justify intervention

The GPAI project is oriented toward a better empirical understanding of how recommender systems work inside a candidate company. The precise relationship being investigated is the tendency for recommender systems to target content to users and distribute it to them. The project anticipates that, if the hypothesised relationship it aims to investigate exists, then this may be harmful, and regulators may wish to intervene. In adopting that position, it seems the GPAI project must inevitably adopt the

[44] Bernstein J, 'Bad News: Selling the Story of Disinformation' (2021) September 2021 Harper's Magazine <https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/> accessed 6 September 2021

position that bare distribution of content – for example, by a recommender system – whether or not it is consumed, and whether or not it has a demonstrable impact on a user, may reach a threshold of harm that justifies regulatory interest.

We raise this because it is important for answering the following question: if we do find that recommender systems produce, in some users, exposure to content that influences them to view similar content that is slightly more "extreme" each time, but is never unlawful or a breach of moderation standards, what then?[45] It is not even clear what why extremeness or radicalness is relevant.

As explained at item 8, there are few categories of content where law or public policy dictates that the bare distribution of such content is unlawful. TVEC and child sexual abuse material are two obvious examples. However, the GPAI project will only encounter such content on the platform if that content has been missed by algorithmic detection systems. The companies routinely remove and prevent this content wherever they are aware of it. They are exposed to legal, financial and regulatory risk if it is found that their systems are missing this illegal content.

Of the content that remains, it is by definition not illegal and not contrary to platform guidelines. The GPAI project is premised on the idea that the distribution of content which is not illegal, and not contrary to platform community guidelines, may potentially be worthy of regulatory attention. This strikes to the core of the debate about whether there is any legitimate interest in States or others hampering the distribution of – not just "awful but lawful" content – but content which may not even be awful. This is likely to have public policy and reputational implications for the GPAI project and for any platform collaborator.

Item 10: The researchers' findings will have serious impacts on the platforms in a range of regulatory areas

If the researchers were to find that the recommender systems have the effect they are testing, then this finding will have direct impacts on the way the platforms are perceived reputationally. This knowledge may also create legal risks for them, insofar as they are perceived to have knowledge that their systems had this effect. Finally, the findings will be relied upon by policymakers and politicians seeking to justify regulatory intervention.

Transparency and content moderation regulation has impacts on other policy areas

The matter of how the social media companies should be regulated and what impacts they have on matters relevant to public policy is very broad. It can frequently be difficult to extricate a particular policy area from another, especially when it comes to the core issue of content moderation. Here are some examples.

1. Concerns about how the platforms moderate content are closely linked to the significant impact that platform moderation decisions have, given the size and dominance of the platforms in global markets. Platform content moderation behaviour is therefore closely linked to potential competition and antitrust intervention.
2. Another class of regulation that is emerging relates to transparency in advertising, which is also linked to concerns about electoral interference and platform manipulation. When the platforms' systems permit certain kinds of ads or other content to be shown to users, this is a form of content moderation. The effectiveness of the platforms detection systems and the

[45] This is not to say that some users, perhaps with pre-existing dispositions, might suffer real harm from increased exposure to content that does not violate the platforms' terms of service, and is not illegal: examples might be children or people who can reasonably be inferred by the platforms to have particular sensitivities.

tendencies of their recommender systems may therefore engage “honest ads” style regulation.

3. The European Union has proposed an Act to regulate the use of high risk artificial intelligence systems. It is highly likely that recommender systems will be caught by this legislation. The researchers’ findings may therefore contribute to the regulation of algorithmic systems which may have wider effects on platform operations.
4. Recommender systems target content based on what the platforms know about an identifiable user, or based on composite profiles of users. As a result, scrutiny of recommender systems is likely to engage questions about privacy concerns, and have consequent impacts on existing and potential privacy regulation.

When it comes to making the case that the companies should be willing to disclose a greater amount of information about their operations, these interrelationships between different regulatory areas is important. It means that a disclosure about how content is moderated could conceivably generate evidence for antitrust action, or claims about bias in algorithmic systems, or inconsistent application of anti-foreign interference policies. It is therefore difficult to argue that the investigation proposed by GPAI will have no impact on a range of other regulatory interventions which the platforms are currently facing.

Item 11: Within the companies’ systems, the recommender systems do not operate in a vacuum.

This item is closely linked to item 1 (Item 1: The project must state the specific concerns and the precise relationship to be investigated). It also links the first class of issues (First class of issues: adopting TVEC as a subject) with the second class of issues (Second class of issues: issues arising from enhanced transparency, whether collaborative or mandated).

With a slightly different emphasis to item 10 (Item 10: The researchers’ findings will have serious impacts on the platforms in a range of regulatory areas), the important point here is that recommender systems are situated within a wide network of algorithmic systems and operational policies and procedures.[46] For that reason, even if the researchers could demonstrate a connection between recommender systems and the proliferation of the content they are trying to assess,[47] the platforms will argue that the impact of the recommender system’s tendencies is minimal. Another point to consider here is that this also creates a risk that (pursuant to item 10) the researchers’ findings are taken out of context and create unjustifiable public perception that creates legal or political risk for the platforms.

This issue is linked to item 1 because it affects the ability of the researchers and of any external party to make a request for access or disclosure that is sufficiently targeted, which therefore raises the principled and practical issues we identify in our discussion of item 1. It also creates challenges for any transparency regime imposed by law, which we discuss below (See: Practical issues facing enhanced access and transparency regimes). In advance of that discussion, we briefly note two points.

Content moderation systems include operational procedures and human decision-making

Another important point to note is that, if recommender systems are content moderation systems, then these content moderation systems are not just technical in nature. They also include a range of policies and procedures that govern how human employees or contractors operate those systems. As

[46] Two examples would be: the content interventions that occur in response to automated and manual flagging of content; and “ad-targeting” or user profiling systems which are used to inform the recommendations to a user. The latter in particular is likely to be extremely commercially sensitive.
[47] Note we do not call it harmful content in light of points 3 and 9.

a result, the kinds of information that illustrate how a recommender system operates must include companies' internal policies and procedures, decision-making documents, guidelines, and perhaps even legal documentation dictating how employees must follow such policies and processes. Again, this is one issue we foresee with attempting to study recommender systems in isolation from the broader socio-technical assemblages that comprise not just the "platforms" as narrowly understood, but the wider companies that operates those platforms.

Recommender systems are interconnected with systems that provide information about user preferences

Recommender systems rely in some part on data inputs about individuals as users in order to provide personalised recommendations. On that basis, it is difficult to extract a recommender system from the wider systems that track and profile users. Any attempt to study recommender systems are therefore likely to expand to considerations of how these tracking and profiling systems operate. This is likely to engage privacy concerns. It is also likely to engage financial and reputational considerations, to the extent that these user profiling and targeting systems may not be as effective as companies' marketing claims them to be.

Item 12: Issues arising from taking a collaborative approach with a social media company

The final challenge we foresee draws all of the previous items together. What we have done is raised a range of legal, financial, regulatory and reputational risks that might be raised by GPAI's proposed approach. We have emphasised that many of these may be able to be resolved, but nevertheless, they will require dedicated work to resolve them.

Some of these risks may not be able to be resolved. Further, in the arena of politics and public relations there is also a risk that even if the issues have been adequately resolved, there nevertheless follows a public outcry, even if unjustified.

Companies are likely to be required to have veto rights

It is highly likely that the GPAI project does create risk to the companies, even where they participate through cautious collaboration. From a structural perspective, given the risks created by collaborative research, it is difficult to see how the platforms could fail to include a right of veto or restriction over what is published. The platforms and their employees have obligations to protect the platforms' rights and interests, as well as the rights and interests of related parties. This means a right of veto is likely to be required regardless of questions of good faith or intent.

Given the risks canvassed in items 1-11, if the findings are harmful to the companies' interests, they are unlikely to be published. If the findings are positive, they will be published. In this way, the existence of a veto right may undermine the perceived integrity of the research regardless of whether such veto rights have any actual impact. Notably, it may even lead to questions about the transparency of the project itself, which might be awkward given the project's emphasis on the value of scrutiny and transparency.

We emphasise that any right of veto held by the companies need not be substantively justified in order to prevent publication or to substantially influence reporting of results: all it needs to do is create enough of a reasonable argument that dispute resolution mechanisms are required to resolve it. Any dispute resolution process will impose burdens of cost and delay that may be significant enough that even unmeritorious objections will prevent the research from being published.

This leads us to make another observation. Many of the legal issues that might arise for GPAI or for other researchers collaborative work with the companies are essentially neutralised by the decision to embed a researcher with the candidate company and to allow restrictions on what that researcher can disclose. Essentially, it risks resolving all the legal barriers to collaboration by simply ceding control of those issues to the companies.

Challenges of taking a collaborative approach illustrate the challenges facing transparency legislation

There is substantial benefit, however, to GPAI's attempt to identify the kinds of issues created by pursuing these collaborative approaches. The benefit is that it allows clear assessment of the specific legal, financial, and reputational barriers to increasing transparency. These insights will be crucial for the forthcoming legislative programmes in jurisdictions like the EU, the UK and the US where transparency and reporting regimes are being explored.

Conclusion (First class of issues: adopting TVEC as a subject)

In combination, we conclude that proposals to study the relationship between recommender systems and TVEC raise challenging issues of law and public policy, including related issues of public confidence and perception. We have raised these issues in order to support the researchers to incorporate them into research design and we have also tried to describe them in ways that will assist policy-makers with an interest in pursuing this kind of research or creating legal frameworks for supporting it.

Having explained the challenges we foresee arising from the subject-matter of the GPAI proposal, we next cover what we have labelled the second class of issues. These relate to the matters of law and public policy that might arise from the decision to study recommender systems inside companies, which engage related questions of how access and transparency around social media systems might be approached. These issues generate important insights too for any future attempts to implement legislation that compels the companies to provide access to researchers or to others seeking to examine the kinds of questions raised by the GPAI researchers.

SECOND CLASS OF ISSUES: ISSUES ARISING FROM ENHANCED TRANSPARENCY, WHETHER COLLABORATIVE OR MANDATED

Overview of this part

To this point we have primarily focused on the legal, reputational, and public policy issues we foresee arising in connection with the subject matter of the proposed research (TVEC). Items 1-12 identify various reasons why the companies might find access and transparency arrangements challenging

The key takeaway from that analysis is that **there are material disincentives, risks, and barriers to companies facilitating access by external parties to their systems, data and other internal materials for research purposes**. We also noted **practical issues exist** whereby it may not be possible to adequately constrain a request for access in ways that do not lead to disproportionate or unjustified disclosures about related systems.

At present, in light of those issues, the current state of affairs is that a broad range of external parties have concluded that:

1. The current state of access and transparency around the companies' operations is inadequate.
2. Superior access should be granted. Presumably, this access is likely to be accompanied by enhanced rights of disclosure too by any external party who successfully achieves access.

The current state of affairs is shaped by an existing network of regulatory restrictions and incentives. If enhanced transparency is to be achieved, then legislation will be required to reconfigure the existing regulatory environment. In this part, we give an overview of the kinds of legal issues that are likely to be engaged by any legislative proposal to impose transparency and access obligations.

In the same way as we set out at item 1,[48] we think there are both practical[49] and principled[50] reasons why transparency is difficult to achieve and why legislation implementing transparency arrangements will face similar issues.

We suggest that a good starting point for policy makers pursuing enhanced transparency arrangements is to assess the insights gathered from previous attempts at voluntary collaborative engagement with social media companies outside of any legislative compulsion. We briefly refer to some of these collaborative arrangements by way of illustration.

We also suggest that the reports produced by the social media companies, whether unilaterally or in partnership with others, provide a useful resource for regulators. These reports are likely to include statements of principle by the companies on why transparency disclosures are desirable. These statements form a body of principle that we can reasonably assume the companies continue to support. Regulatory arrangements that are framed in such ways are therefore less likely to meet with vigorous opposition.

In pursuing a legislative reconfiguration of existing rights and obligations, States should adopt a human rights approach. By a human rights approach, we mean an approach which starts with a consideration of the relevant human rights which are engaged by a particular research or regulatory proposal. When it comes to imposing regulation which might limit human rights, any limitations should adhere to longstanding principles of legality, necessity and proportionality.

---

[48] Item 1: The project must state the specific concerns and the precise relationship to be investigated.
[49] Practical reasons for specificity.
[50] Principled reasons for specificity.

A human rights approach can help to generate consensus even among opposing stakeholder groups as a minimum statement of principle. Platform companies have also, to varying degrees, committed to respecting human rights in line with the UN Guiding Principles on Business and Human Rights, and investor groups oriented toward responsible investment practices have commensurate expectations. Human rights approaches do not necessarily need to be framed as such in jurisdictions where there is hostility to human rights instruments, or they have lesser importance, for example in the US. The key thing is that a human rights approach allows for close attention to the various rights at stake. This remains important even in western liberal democracies. We note that across jurisdictions there is a real risk that transparency approaches might incidentally confer indirect powers on State-level actors in ways that were not intended[51] and a cautious approach is important, despite the urgency of calls to action.

For the benefit of policymakers and for researchers pursuing collaborative approaches like GPAI, we give a brief summary of the key features of selected legislative proposals. These proposals are useful because they are comprised of specific statements about precisely what regulators are proposing will be required of the companies in order to implement transparency-oriented policy approaches. Stating these obligations in the form of a regulatory proposal encourages more specific points of agreement or disagreement among stakeholders with a view to meaningful progress. However, as will be seen, the specific legislative proposals that are currently proposing the implementation of transparency arrangements are largely silent on the crucial details that we perceive will provide practical and principled barriers to their implementation.

There are real challenges for the companies created by enhanced access and transparency arrangements

In our discussion of items 1-12 above, we have raised a range of legal, reputational, financial and regulatory issues that could arise if the companies collaborate with external parties to permit greater access to their internal systems. These issues are important because it is easy to assume that transparency and access arrangements for trusted parties (whether researchers or regulators) pose little risk to the companies. A related assumption is that any risk to the companies from greater transparency will necessarily be justified, in the sense that the companies have nothing to hide if there is no real cause for concern, and any resistance can only be a kind of cover-up.

In practice, it is highly likely that many of these issues have been navigated before through previous attempts at collaborative research engagements. We describe some of these below and suggest that closer analysis of them by policymakers would be useful.

Policy makers should closely examine existing collaboration and transparency arrangements

We think policy-makers should start by close examination of existing initiatives directed toward enhanced transparency, access and collaboration. We briefly summarise notable examples below and organise them according to whether they are unilateral or multilateral, and whether or not they include States. The latter is an important factor given the human rights risks that could arise from providing inadequately fettered access to social media data, systems, and operational processes.

The purpose of any future analysis by policymakers would be to establish:

---

[51] See the comment by advocacy body Article 19 on the potential risks of art 27 of the DSA. ‘At a Glance: Does the EU Digital Services Act Protect Freedom of Expression?’ (ARTICLE 19) <https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/> accessed 8 September 2021.

1. What are the current or historic legal and extra-legal barriers to enhanced access and transparency for external third parties, whether researchers or otherwise.
2. When regulating for transparency, which of those identified barriers will be addressed, including by altering rights and obligations held by existing legal and natural persons.

We emphasise that our decision to refer to a range of collaborative initiatives below does not carry with it any implied approval or endorsement as to the effectiveness of these approaches. By contrast, it is highly likely that the people engaging with those approaches (whether unilateral or collaborative) will have many specific and tangible suggestions about ways that they could be improved. It is these insights which ought to be incorporated into any policy process directed at implementing legislative arrangements to enforce transparency and access for external parties.

We note that our summary of relevant initiatives is skewed toward initiatives directed toward counterterrorism. This is partly because of the focus of the GPAI project and Brainbox's previous work in this area. By contrast, many of these initiatives arose at a time of escalating State-level concerns about lone wolf terrorist attacks, the distribution of terrorist content, and potential pathways toward radicalisation on social media.

Examples of existing collaborative arrangements

Some collaborative arrangements consist of **partnership between the companies and external research institutions**, particularly universities or research centres within universities. Notable examples include:

- Social Science One: this was a wide-ranging stakeholder collaboration with significant partnerships between tertiary institutions and the companies. This programme faced significant issues, including unforeseen delays in providing access to desired data. Further, it was recently discovered that of the limited data that could be provided, that data was unreliable, and required subsequent correction. Professor Nate Persily has drafted proposed transparency legislation for the United States, and he was co-Chair of Social Science One for a period.
- Assistant Professor Kate Klonick is a professor of law at St John's University. She conducted a programme of work closely observing the set up and initial operation of Facebook's Oversight Board proposal. Academics who have entered into limited research partnerships with the platforms have achieved a higher level of access and transparency than is otherwise available to the public or to other researchers, and in this sense have entered into a collaborative transparency arrangement.
- We also note that there are academics who have direct experience of working at relevant platform companies, such as Daphne Keller, who now directs the Program on Platform Regulation at Stanford's Cyber Policy Centre. Researchers with direct experience of the companies' systems bring their own kind of transparency and access, given the superior knowledge they have of how company systems are configured and the kinds of factors that contribute to company decision-making.

There are a number of collaborative transparency approaches that have been adopted by one or more parties. Some of these collaborative arrangements **include nation states, or nation-state level institutions**. Arrangements of this kind include:

- Tech Against Terrorism: a partnership between tech companies and the United Nations Counter Terrorism Executive Directorate. One notable output from this collaboration is the Knowledge Sharing Platform, intended to be used by smaller companies struggling to manage terrorist content with more limited resources than the larger platforms. In 2018, it launched "The Data Science Network" which is a "network of experts working on developing and deploying automated solutions to counter terrorist use of smaller tech platforms whilst

respecting human rights."[52] The organisation also produces a range of research outputs about key trends in TVEC. It also provides guidance about how to perform transparency reporting, best practice policy and a range of other resources.

- The Christchurch Call is a voluntary partnership between tech companies and nation states. It is premised on respect for human rights and a free and open internet. Countries and tech platforms can join the call by pledging to respect a set of values. The Christchurch Call has a civil society advisory group that has expressed concerns about human rights impacts of the Call. There is also an emerging issue about how the Call and its supporters should respond if a signatory is perceived not to have honoured its commitment to the Call's values.
- The European Union's Code of Conduct on Countering Illegal Hate Speech Online: since 2016, the EU has conducted regular auditing and monitoring exercises with key technology companies to monitor compliance with the EU Code of Conduct on Countering Illegal Hate Speech Online. The audit process is confidential, but the results of the audit are publicly available. "Since adoption in 2016, the Code of Conduct is delivering continuous progress: the last evaluation shows that on average the companies are now assessing 90% of flagged content within 24 hours and 71% of the content deemed illegal hate speech is removed."[53]

A human rights approach to the internet acknowledges that States pose unique risks to the internet given the power and authority that is only available to Nation States, for example, a monopoly on the lawful use of violence and powers of detention and taxation. Given the risk that States can pose to the human rights of platform users, **some multilateral transparency approaches do not include States** within their membership. These organisations include:

- Global Network Initiative: this is a partnership of technology companies, academic organisations, academics, civil society, investors, observers, and fellows. Members commit to a set of principles oriented toward human rights to freedom of expression and privacy. It is committed to transparency and accountability and multi-stakeholder collaboration. The current independent Chair of the GNI Board is David Kaye, who was until recently the UN Special Rapporteur for Freedom of Expression.
- The Global Internet Forum to Counter Terrorism ("GIFCT"): this body was set up in response to rising concerns in connection with terror attacks thought to have been influenced by radicalisation on social media. For much of its history it had no independent status from its founding members, however in the aftermath of the Christchurch attacks, its transition to a separate body was accelerated. The GIFCT is primarily composed of technology companies and its membership recently expanded to 17 companies. The GIFCT has partnership with a separate research network, "GNET", which has produced a body of research, although it is unclear whether this research benefits from enhanced access and transparency with social media companies. The GIFCT has responsibility for the shared hash database and the Content Incident Protocol for responding to live terror events with a substantial likelihood of a viral online component.
- Oversight Board (Facebook): Facebook recently established a separate institution funded from an endowment and operated under an independent charter known as the Oversight Board. The Board is comprised of a range of notable individuals from academia, law, government, and civil society. The Board has called numerous times for enhanced information from Facebook and broadly been unsuccessful. In particular, it is scheduled to speak with Frances Haugen, specifically in light of revelations about the "X Check" programme. Haugen has alleged that Facebook has lied to the Oversight Board repeatedly.

---

[52] 'About Tech Against Terrorism - Tech Against Terrorism' (4 September 2017) <https://www.techagainstterrorism.org/about/, https://www.techagainstterrorism.org/about/> accessed 12 October 2021

[53] 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (European Commission - European Commission) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 1 April 2021

The Oversight Board recommended an analysis of how Facebook's systems might have contributed to a narrative that undermined the legitimacy of the 2020 US Presidential election.

The social media companies have also adopted various practices that we would describe as **unilateral approaches to transparency**. These consist primarily of "transparency reports" that are prepared and released by the companies. Again, referring to these reports should not be taken as an implied suggestion that they are sufficient, however they do provide a body of principle and insight as to how transparency reporting currently operates and the policy basis for conducting such reporting. Examples include:

- The biggest social media companies conduct generic periodic transparency reporting. Some of this reporting includes machine readable data sets.
- Occasionally, social media companies will release more fulsome transparency reports on particular issues. These reports are frequently used to deal with issues of high public significance, including action against networks of users perceived to have been acting with coordinated intent to influence public perception, especially by States.
- Under this category we would also include the platforms' own substantial research initiatives, from which a range of publications are produced. Again, this is not to suggest that these publications by themselves are sufficient, only to suggest that the do provide some insight into the existing features that influence the disclosure of information about platform operations.

We also note that, from time to time, there are ad hoc statements of principle or collaborations from civil society. These do not always involve collaboration or partnership with the platforms. Under this category we would include:

- the Santa Clara Principles on Transparency and Accountability in Content Moderation[54] and
- the Manila Principles on Intermediary Liability.[55]
- We would also include broader academic work on matters engaged by transparency related concerns under this category.[56]

Finally, we also note that transparency and access has been facilitated through non-collaborative means, specifically in the context of **leaks, whistleblowing, and law enforcement actions.** Again, by studying how incidents of this kind have occurred, and the kinds of legal issues that arose, we will gain superior insight into the way that mandated transparency and access regimes will need to operate and the kinds of legal barriers they might face.

Selected legal issues likely to arise

Broadly speaking, we think it is clear the following kinds of legal issues are likely to arise in most jurisdictions. The extent of them will obviously vary depending on the specifics of the transparency arrangement being considered[57] and the domestic legal instruments involved.

- Potential risks to user privacy from access to and potential disclosure of data.
- Intellectual property risks from disclosing the details of how algorithmic systems operate.
- Commercial confidentiality considerations from disclosing how business systems operate, including wider operational processes and human decision-making processes.
- Commercial considerations might arise where external parties perceive that platforms have made misrepresentations about their systems to customers, including advertisers, whether

[54] https://santaclaraprinciples.org/
[55] https://manilaprinciples.org/principles.html
[56] As one example, see: Suzor NP and others, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communication 18.
[57] We refer again to our discussion at Item 1: The project must state the specific concerns and the precise relationship to be investigated

related to the reach of advertisements, or other matters such as “brand safety” (referring to advertisers’ intolerance for having their brand displayed next to particular kinds of content).
- Competition concerns related to competition between the companies and potential insights a competitor may gain into another competitor’s systems.
- Compliance burdens: depending on the company involved and its existing systems for collating reports about how its systems operate, there may be significant resourcing required to comply with transparency legislation. Notably, the DSA imposes differential obligations depending on the size of a company, and organisations such as Tech Against Terrorism have worked to support smaller companies to deal with the issues the larger companies face around TVEC.
- Disclosure of system features might lead to assurance and integrity issues related to potential adversarial or antagonistic engagement with those systems once details are known.
- There is a risk that selected disclosures about one class of systems leads to unintended disclosures about other classes of systems. What legal arrangements should exist against disclosure or investigation of incidental systems?
- The law will have to consider the rights of third parties who are neither users nor employees of social media products, but might be contractors or consultants who perform operational functions.[58]
- Transparency approaches may touch upon the rights of customers (advertisers) or publishers and content distributors (for example, note the disclosures about the impact on Buzzfeed of algorithmic changes at Facebook in the disclosures by Frances Haugen).

These legal issues should not be conflated with the various human rights that might also be engaged by the platform companies’ conduct. The list above refers to sources of existing legal protection that cement the current regulatory arrangements. Separately, the task of regulators is to consider how far these protections or obligations ought to be restricted or enhanced. In making these assessments, regulators will be required to consider the following questions.

- What is the strength of the evidence showing that the platforms’ products cause harm or merit independent scrutiny? This is a specialist question that ought not to be assessed purely on the basis of second-hand summaries of research results, for example in the news media.
- Is the alleged harm of a kind recognised by human rights instruments? Even if harmful, is the harm necessary in free and democratic societies? For example, despite their potential harms, freedom of expression about public figures, or freedom of opinion on reasonably contested political issues may not be able to be legitimately prevented.
- How will the proposed transparency arrangements actually contribute to mitigating the alleged harms? How is transparency going to help?
- What trade-offs might be involved by trying to mitigate these harms? Are these trade-offs acceptable?
- Importantly, even if action is justified, what safeguards and accountability measures are incorporated into legislation to prevent abuse of the powers it creates? The legislation should not assume that government intentions will always be good: this is a core feature of human rights instruments and the use of law to constrain the power of the state. Legislation should confer rights to a remedy in case of abuse, or rights of appeal for affected parties in case a decision or approach has occurred on the basis of an error of fact or law.

Practical issues facing enhanced access and transparency regimes

We also emphasise our concern that it may be difficult to study individual systems (like recommender systems) in isolation, without also triggering consequent investigations into a wide array of

[58] Consider for example the rights and interests of external contractors like Accenture or other smaller content moderation sub-contracting firms. See ‘How Facebook Relies on Accenture to Scrub Toxic Content - The New York Times’ <https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html> accessed 6 September 2021.

interconnected systems. The GPAI focus is on recommender systems and describes a method for precisely isolating the causal impacts of these systems. However, when it comes to describing (in legislation or by access requests) which systems researchers want to study, we anticipate conceptual and practical difficulties when attempting to disentangle these systems.

The DSA adopts definitions of recommender systems and content moderation systems. A recommender system is defined as follows:

> *(o) 'recommender system' means a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service, including as a result of a search initiated by the recipient or otherwise determining the relative order or prominence of information displayed;*

Content moderation systems are defined as follows:

> *(p) 'content moderation' means the activities undertaken by providers of intermediary services aimed at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility and accessibility of that illegal content or that information, such as demotion, disabling of access to, or removal thereof, or the recipients' ability to provide that information, such as the termination or suspension of a recipient's account;*

This conceptual distinction is useful for imposing specific obligations on recommender systems that might not make sense when describing content moderation systems. For example, the DSA imposes obligations on very large online platforms to disclose the parameters used in recommender systems. However, when it comes to subsequent risk assessments to be conducted by the DSA, these relate to "how their content moderation systems, recommender systems and systems for selecting and displaying advertisement" contribute to systemic risks.[59] Further, when it comes to imposing mitigation measures to guard against these systemic risks, mitigations may be applied to:[60]

> *(a) adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services, or their terms and conditions;*
>
> *(b) targeted measures aimed at limiting the display of advertisements in association with the service they provide;*
>
> *(c) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk;*
>
> *(d) initiating or adjusting cooperation with trusted flaggers in accordance with Article 19;*
>
> *(e) initiating or adjusting cooperation with other online platforms through the codes of conduct and the crisis protocols referred to in Article 35 and 37 respectively.*

The point is that the impact of a recommender system may be very difficult to assess in isolation from a range of other technical and non-technical systems. As Llanso et al state, recommendation systems are already a form of content moderation.[61]

> *As discussed recommendation systems are not inherently neutral and are designed to prioritize content with certain characteristics and deprioritize other content. As such, their functioning is already a form of content moderation: by suggesting some types of content and hiding others, they perform an important gatekeeping function.*

Grimmelman defines moderation broadly as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" and notes that these mechanisms

---

[59] Article 26.
[60] Article 27
[61] Llansó E and others, 'Artificial Intelligence, Content Moderation and Freedom of Expression' (Transatlantic Working Group 2020) at p 19.

themselves can be automated, such as recommender systems.[62] Equally, content moderation systems themselves are not restricted to computational systems. At a recent discussion hosted by GIFCT, Brian Fishman (Facebook) referred to the human coordination effort required to update the way 35,000 human content moderators operate when systems change. In relation to the expanded taxonomy for the hash database, he said:[63]

> *... it's sort of the classic cliche around turning an aircraft carrier. You've got a big operation - not just of algorithms and of automated systems that remove things, but most importantly it starts with human reviewers - there are 35,000 human reviewers reviewing content at Facebook and building new decision trees for those folks, training all of those folks, making sure that you've got the right quality in place so that you can make good decisions and especially when you're asking for more nuance ... for the academics in the room, I think this speaks to the biggest learning that I have coming into a large tech company is that … the sort of precise narrow distinctions that you want to make all the time are very hard to make at the scale that Facebook is operating, because there are so many reviewers. That many reviewers with that kind of diversity among those reviewers - the great advantage is that you've got this cultural knowledge, linguistic knowledge that spreads all over the world. Folks can really dig into lots of different problems that no individual or small team or small team of experts is able to wrap their head around. The challenge is that many of those folks are not experts in these issues. They don't have the conceptual background. They haven't all taken a masters or a PhD in studying political violence and terrorism and getting them to the point where they can consistently make decisions that reflect nuanced policy choices is really hard.*

We think a useful way of thinking about this is summarised by Gillespie, who attempts to articulate a definition of a "platform" and concludes that content moderation is a constitutional activity for the companies:[64]

> *To the definition of platforms, I would like this book to add a fourth element: d) platforms do, and must, moderate the content and activity of users, using some logistics of detection, review, and enforcement. Moderation is not an ancillary aspect of what platforms do. It is essential, constitutional, definitional. Not only can platforms not survive without moderation, they are not platforms without it. Moderation is there from the beginning, and always; yet it must be largely disavowed, hidden, in part to maintain the illusion of an open platform and in part to avoid legal and cultural responsibility. Platforms face what may be an irreconcilable contradiction: they are represented as mere conduits and they are premised on making choices for what users see and say. Looking at moderation in this way should shift our view of what social media platforms really do: from transmitting what we post, to constituting what we see. There is no position of impartiality. Platform moderators pick and choose all the time, in all sorts of ways. Excluding porn or threats or violence or terrorism is just one way platforms constitute the social media product they are generating for the audience. The persistent belief that platforms are open, impartial, and unregulated is an odd one, considering that everything on a platform is designed and orchestrated.*

There is a real risk that these conceptual, technical, and other issues mean the platforms are incapable of complying with requests for information, even in extremely well-resourced initiatives including diverse stakeholders. We have no specific insight into whether the issues we raise here are relevant, but the funders of the Social Science One project ultimately concluded that:[65]

---

[62] Grimmelmann J, 'The Virtues of Moderation' (2015) 17 Yale J.L. & Tech. 42.
[63] '2021 Global Summit' (GIFCT) <https://gifct.org/2021-global-summit/> accessed 6 September 2021 – session 2 https://vimeo.com/581218164 - Brian Fishman at 18m et seq. Transcribed by Brainbox.
[64] Gillespie T, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (2018) at p 21, available from: <https://www.researchgate.net/publication/327186182_Custodians_of_the_internet_Platforms_content_moderation_and_the_hidden_decisions_that_shape_social_media>.
[65] 'Funders Are Ready To Pull Out Of Facebook's Academic Data Sharing Project' (BuzzFeed News) <https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data> accessed 8 September 2021

> *the technical and legal complexities associated with making proprietary data available to independent scholars are greater than any of the parties originally understood, and Facebook has as a result been unable to deliver all the data initially anticipated. The 83 independent scholars whose proposals were selected for funding have access to only a portion of what they were told they could expect, and this has made it difficult or, in some cases, impossible for them to complete the approved research. Nor can Facebook or its privacy and security advisory committees yet offer a definitive timetable for when the full set of proposed data can be made available.*

The difficulty of extracting content moderation from content recommendation is also illustrated by this description of the process by MacCarthy, where the removal of infringing content from the pool available to recommendation algorithms is a necessary part of their operation:[66]

> *Platforms typically screen all content via matching algorithms to reidentify content which has already been identified as inappropriate and fingerprinted in a database for that purpose. For instance, platforms consult their own fingerprinted databases of content that previously was found to be terrorist material or child exploitation, and check hashtag databases maintained by external organizations such as GIFCT for terrorist material and the Internet Watch Foundation for child exploitation images. After this initial screening, the material is posted and subsequently subjected to further automated screening using systems deemed sufficiently reliable to determine likely violation of standards. If the automated system shows a clear violation, such as material that is highly likely to be nudity or a new child exploitation image or fresh terrorist content, the material is removed without further human review. If the judgment is uncertain, or if the material has been flagged by a user as a potential violation, then it is routed to a human reviewer. Platforms sometimes sample reviewer decisions and subject them to further review to determine the "correct" decision.*

If recommender systems and content moderation systems are inseparable, then there is a real risk that systems for classifying content and for profiling users in order to deliver content and advertising are also inseparable. Increasing the scrutiny toward user profiling and content targeting could create financial risk to the companies. It is possible that there is a degree of marketing "puffery" around, for example, the size or make-up of a company's audiences, or the effectiveness of its products and systems. The capabilities of the companies' systems may have been overstated even if there has been no culpable misrepresentation amounting to fraud.[67] By complying with requests about one system, the companies may therefore risk revealing details about other systems, leading to significant legal, reputational, and financial impacts even if only minor exaggeration of claims has occurred. When it comes to market responses, even perceptions of this kind will suffice, even if unjustified.

Clearly, for practical purposes and at a conceptual level, the companies will segment the way that they administrate these systems. In fact, even the ability to enumerate and describe the systems above does suggest a capacity to distinguish between them. However, that ability to focus on conceptually distinct systems operationally does not mean that the impact of each of these systems can be meaningfully studied in isolation. That is because the impacts of these systems are likely to be interrelated. That leads to the prospect that studying one system may lead to a daisy-chain effect where greater and greater amounts of information is disclosed.

This issue of whether requests can be sufficiently tailored also has ramifications for any legal regime that aims to foster transparency arrangements. There is a risk that any legal mandate for obtaining information will be inconsistent with the human rights requirements of proportionality and necessity. The real world effect of this will be that:

---

[66] MacCarthy M, 'Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry' (Transatlantic Working Group 2020)

[67] Consider for example the claims made in response to the following kinds of analyses: 'Did Facebook's Faulty Data Push News Publishers to Make Terrible Decisions on Video?' (Nieman Lab) <https://www.niemanlab.org/2018/10/did-facebooks-faulty-data-push-news-publishers-to-make-terrible-decisions-on-video/> accessed 12 October 2021

1. It may not be possible to adequately circumscribe a request for information such that it is proportionate and necessary. Requests may “snowball” or become impossible to comply with, or lead to obligations to disclose material in ways that are not tenable from a public policy perspective.
2. Relatedly, this issues around the interconnectedness of systems may undermine the ability of regulators or researchers to be able to make persuasive claims about the impact of social media systems. For example, companies may always be able to point to another system that neutralises any adverse impact of the system under investigation. The ability to use the insights gained from transparency arrangements may therefore be limited.

A human rights approach allows for a principled analysis of the web of rights and obligations

A principled approach to limiting human rights

At a principled level, any regulatory regime will impose access requirements in ways that inevitably impose restrictions on existing rights and interests. To the extent a human rights approach to these restrictions is to be followed, the restrictions should meet the following broad criteria.[68]

- The restriction (the obligation to provide access) should be imposed by law. It should be capable of being reasonably clear and comprehensible. A company should be able to voluntarily comply in good faith. If the legal instrument being imposed on the companies is too vague or discretionary, then it risks becoming an arbitrary enforcement tool that can be used by States for selective and arbitrary enforcement.
- The restriction should be proportional to the harms involved and linked to a legitimate regulatory objective. For example, it would not be legitimate to impose transparency obligations that could hamper user freedom of expression if the alleged harm relates to identifying anonymous critics who are making true statements about democratically elected public officials. Further, to impose broad and non-specific transparency obligations could lead to onerous transparency obligations out of all proportion to the harms involved.
- In close association with obligations of legitimacy and proportionality, the restrictions imposed on the existing rights and obligations held by companies, users, employees, shareholders and directors must be necessary for achieving the specified regulatory objective. There must be a real link between the restrictions being imposed and the harms being caused, otherwise States could broadly assert there is a link between non-specific conduct and a vague harm and impose restrictions entirely unconnected with the conduct involved. Human rights bodies have concluded there must be a demonstrable connection between the restriction being imposed and the regulatory objective being pursued.

At a principled level, it is not possible to conduct informed democratic debates that weigh and balance competing interests and trade-offs if we have not been adequately specific about what we are proposing to achieve and what we are proposing to alter in order to achieve that. For example, if we fail to recognise that transparency obligations are likely to impact on the rights of contractors or employees, then we are unable to account for these as part of the policy process.

Another point to consider when it comes to weighing and balancing various human rights is that we must be clear about the breadth of human rights involved.

- A curious point to consider is the way that international human rights instruments (including the Universal Declaration) do confer some significance on rights to property (owned

[68] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35, Human Rights Council, Thirty-eighth session (18 June – 16 July 2018). See also Universal Declaration of Human Rights, art 29(2).

individually or jointly with others),[69] the right to be protected by the law from unjustified attacks on honour and reputation,[70] and rights to intellectual property.[71] These rights probably sit on the side of the companies rather than the side of users and States and weigh against transparency arrangements. This is not to say those rights are absolute or to express a personal view on how they should be weighed against other rights. Article 29 of the Universal Declaration of Human Rights also states that "everyone has duties to the community".

- Against the considerations raised by the rights immediately above are the rights that many allege the platforms are currently infringing. These include rights to personal safety,[72] to privacy,[73] to health,[74] to freedom of conscience, association, opinion and expression,[75] to public participation in the community and democratic processes,[76] and freedom from incitement to violence and discrimination.[77] These are the matters most obviously affected by platforms' products on users.

Our point is that all of these rights are engaged at some level by proposals to impose transparency requirements and the current global discussion about social media products. Regulators cannot credibly weigh and balance them if they fail to realise they are engaged.

It is likely that authoritarian regimes will emulate the worst features of democratic regimes

The United Nations Special Rapporteur for Freedom of Expression has prepared a range of reports on social media regulation.[78] Those reports speak favourably about the role of transparency as a check on human rights abuses. However, we note that transparency is just as often regarded as being a useful means for shedding light on abuses of power by States as it is for platforms. Any transparency and access legislation should include transparency obligations on States to detail how they are using such legislation and preserve the ability of the platforms to push back against unlawful or illegitimate uses of whatever legislative powers are conferred.

Another crucial insight from human rights bodies is the way that legislative action by democratic states committed to human rights will be taken as validation of similar approaches being adopted by states with a more questionable human rights record. In essence, States will respond to criticism from other States by saying, "You do it, so why can't we do it too?"

Existing legislative proposals for transparency

As we have said, we believe most collaborative arrangements are likely to reach a point of limitation which mean States will inevitably transition to legally mandated frameworks for access to social media companies' information. That is because it is almost certain that the companies will attempt to limit the publication of findings if external parties reach conclusions which are adverse to the interests of the companies. We emphasise this does not have to be a result of malevolent intention by the companies: there is ample room for good faith disagreement on a wide range of matters in this developing area of research. Where the companies and the external party disagree, what then? The role of legislation is to anticipate these disagreements and to articulate a way to resolve them.

---

[69] Universal Declaration of Human Rights, art 17.
[70] UNDHR, article 12.
[71] UNDHR, article 27.
[72] UNDHR, arts 3, 4, 5, 6, 7 and others.
[73] UNDHR, art 12.
[74] UNDHR, art 25 and subsequent socio-economic rights instruments.
[75] UNDHR, art 19, 20, 18,
[76] UNDHR, art 21, 27 and 29.
[77] UNDHR, art 7.
[78] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35, Human Rights Council, Thirty-eighth session (18 June – 16 July 2018).

Legislation will enhance legal certainty for the companies, for external parties, and for other observers who wish to assure themselves that research has not been compromised by researchers' desire to maintain access to the platforms' systems.

In this part, we offer some brief comments on key legislative proposals that incorporate a transparency and disclosure arrangement.

European Union: Digital Services Act proposal

The European Union's Digital Services Act creates a range of access and transparency obligations.[79] Key insights for GPAI from the DSA include that:

- The kinds of legal and practical matters we raise in this report have been anticipated in the DSA, including the legal and extra-legal risks and considerations faced by the platforms in exposing access to their systems.
- The DSA defers definition of TVEC to the definitions adopted by the EU and by members states, thereby avoiding the definitional difficulties faced by the GPAI project and by other advocates of content-specific regulation of the platforms.
- The DSA itself illustrates the interrelationship between a range of key systems and the potential that they are difficult to extricate from one another when it comes to assessing their impacts.
- The DSA itself requires the difficulties created by these practical and legal matters to be dealt with collaboratively through subsequent processes.
- The platforms will be anticipating both the implementation of the DSA and their engagement in these subsequent processes. This may create opportunities for GPAI to collaborate with platforms to test future approaches.

TVEC is a form of "illegal content" which is to be assessed by reference to laws of the Union or a member State (art 2(g)). The DSA imposes significant procedural limitations on the way that States can seek action against illegal content or obtain information about identifiable individuals (for example, art 14). Article 21 would impose disclosure and notification obligations on platforms to notify law enforcement if they become "aware of any information giving rise to a suspicion that a serious criminal offence involving a threat to the life or safety of persons has taken place, is taking place or is likely to take place".

As set out above, both "content moderation" and "recommender systems" are defined (article 2, (o) and (p)). Content moderation can include demotion of content (presumably, through altering a recommender system) and recommender systems are defined as fully or partially automated systems for "determining the relative order or prominence of information displayed". Article 29 obliges platforms to set out the parameters used in recommender systems in their terms and conditions "in a clear, accessible and easily comprehensible manner" and any optional parameters available to users. Where optional parameters exist, platforms must create functionality for users to modify these parameters.

A core framework to consider for GPAI relates to the obligation to conduct risk assessments and to mitigate risks in arts 26 and 27. Article 26 requires "significant systemic risks" to be identified, analysed and assessed. The risk assessment "shall include the following systemic risks": dissemination of illegal content (such as TVEC); impacts on a range of specific human rights; and intentional manipulation of platform services, such as through inauthentic behaviour. Article 26(2) requires particular attention to be paid to the following:

> *how their content moderation systems, recommender systems and systems for selecting and displaying advertisement influence any of the systemic risks referred to in paragraph 1, including*

[79] The following analysis relies on the wording as at 14 October 2021 drawn from the following: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0825&from=en>.

> *the potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions.*

As outlined above, the mitigation risks that might be taken pursuant to article 27 illustrate the interaction between a range of social and technical systems, including "adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services, or their terms and conditions;" and "reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk". Article 27 anticipates the publication of "best practice" guidelines and recommended measures for mitigating systemic risks in light of their potential human rights implications. Such guidelines must be subject to public consultations.

Article 28 obliges platforms to submit to – at their own expense – annual audits to assess compliance with specific obligations under the DSA. Such audits could lead to "operational recommendations", and platforms must report on how they implement these.

Of note for the GPAI project, article 31 sets out a framework for "Data access and scrutiny".

- The Commission, or nominated representatives of member states, can seek "access to data that are necessary to monitor and assess compliance with this Regulation" and the platforms must provide this upon "reasoned request and within a reasonable period". Such data may only be used for "those purposes".
- Platforms must also provide "access to data to vetted researchers" who met requirements set out in the DSA. The "sole purpose" the information can be used for is "conducting research that contributes to the identification and understanding of systemic risks as set out in Article 26(1)".
- Access to data under either of the arrangements above (for regulatory assessment of compliance, or vetted researcher analysis of systemic risks) must be provided "through online databases or application programming interfaces, as appropriate".

The requirements for vetted researchers reflect the reality of the trade-offs we have identified in this report. In particular they must:

- Be affiliated with academic institutions and "have proven records of expertise in the fields related to the risks investigated or related research methodologies".
- Be independent from commercial interests
- Commit to, and "be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request".

Delegated legislation will be used for "laying down the technical conditions under which very large online platforms are to share data" for vetted researchers or with regulatory compliance bodies. This delegated legislation will further clarify:

> *the specific conditions under which such sharing of data with vetted researchers can take place in compliance with [the GDPR], taking into account the rights and interests of the very large online platforms and the recipients of the service concerned, including the protection of confidential information, in particular trade secrets, and maintaining the security of their service.*

The platforms may, in response to a request by vetted researchers or by regulators, ask that the request is amended on two grounds and propose an alternative means of access "which are appropriate and sufficient for the purpose of the request". The grounds reflect matters we have raised above and include:

1. That the platform "does not have access to the data".
2. That "giving access to the data will lead to significant vulnerabilities for the security of its service or the protection of confidential information, in particular trade secrets."

As such, the DSA defers all of the complicated trade-offs and practical considerations we raise in this report to a subsequent legislative process, which will be implemented after consultation and in light of

the specific barriers contemplated by various parties. There is little guidance as to how these trade-offs should be managed other than the broad legislative direction provided by arts 26 and 27 that risks should be assessed and mitigated, and that these risks and mitigations could relate to content moderation systems, to recommender systems, or to a broad range of conceivable operational policies and processes.

Notably for GPAI, article 34 anticipates the adoption and implementation of a range of voluntary standards cover submission of data, use of APIs, auditing processes, interoperability considerations and other factors. Codes of conduct will also be produced under arts 35 and 36 for other areas of compliance related to risk assessments and online advertising. It anticipates civil society involvement in formulating these codes.

The wording of the DSA is still subject to negotiations among EU members and may change significantly before it is implemented. The DSA contains a significant enforcement section which itself includes authorities for investigatory bodies to request and receive information and other forms of investigatory behaviour.

Prof Nathaniel Persily's draft

Professor Nathaniel Persily was a co-Chair of the Social Science One research initiative. When Frances Haugen made her disclosures about Facebook, Prof Persily made available on social media a draft of proposed transparency legislation for the United States.[80] Notably, in a panel discussion following Haugen's disclosures, Prof Persily noted that transparency legislation does not have a purely observational purpose, and is also likely to induce specific changes by the platforms, saying: "It's a way of controlling the companies because they know they are being watched …"[81]

Our key conclusion from examining this legislation is:

- That it is primarily oriented toward removing privacy as a barrier to sharing information.
- It defers any specific guidance on how to resolve other related matters, such as intellectual property or commercial confidentiality, to other legal instruments and to delegated regulation.
- It anticipates a broad range of potential issues of the kind we raise above in this report.

The draft is directed toward "qualified data and information", which may include (s 3(8)):

> *information about User exposure, engagement, and other behaviors; data about content producers and content production policies; information that the Qualified Platform otherwise makes available for sale to commercial entities or advertisers; information that goes into the preparation of reports that Qualified Platforms provide to the government or other entities, such as those relating to enforcement of community standards; and metadata related to any of the preceding categories*

It also defines a "qualified researcher" to mean:

> *a university-affiliated researcher conducting research according to a research plan that has been approved by the Division or its appropriate delegate. No employee of a state of federal law enforcement agency or any government employee except for a university-affiliated researcher shall be considered a Qualified Researcher.*

The legislation broadly sets out obligations on platforms and on researchers, as well as immunities in conferred from complying with those obligations. Section 4 states an obligation to comply with relevant laws and an immunity from litigation from releasing "data" in compliance with the enactment. The conditions on providing access to data and information include:

> *(1) encryption of the data in transit and at rest;*

---

[80] See <https://twitter.com/persily/status/1445409819486216203?s=20>.
[81] 'The Facebook Files - Yale Law School' <https://law.yale.edu/isp/events/facebook-files> accessed 8 October 2021

> *(2) delivery of data in a format determined by the Division that is not reasonably capable of being associated or linked with a particular individual;*
>
> *(3) use and monitoring of a secure environment to facilitate delivery of the Qualified Data and Information to Qualified Researchers while protecting against unauthorized use of such data;*
>
> *(4) evaluation by the Qualified Platform of any results garnered by Qualified Researchers before submission for publication but only to prevent public release of Personal Information or other violations of law.*

A notable feature of the legislation is that it simply refers to a range of "federal, state, and local information sharing and privacy laws and regulations" as well as "all rules, standards, regulations, and orders" issued by the FTC "pursuant to this act which are applicable". The legislation therefore confers a considerable burden on the platforms and others to identify these relevant laws and to ensure compliance with them. It is not clear from the face of the legislation itself (and our knowledge of American law) whether platforms may face conflicting obligations to both comply with the new transparency law and comply also with other laws that mitigate against compliance.

Like the DSA, Persily's draft legislation requires a subsequent regulatory process whereby regulations are created to deal with many of the kinds of specific issues raised in this report that might arise from conferring greater access to and disclosure of information.

Researchers are granted an immunity if they are conducting qualified research projects and a division of the FTC is established and authorised to "develop and establish recommended standards, criteria, and approval process for Qualified Researchers, Qualified Research Projects, Qualified Data and Information, and Qualified Platforms" pursuant to consultative processes. The criteria for qualified researchers may not include:

> *consideration of political views, race, gender, gender identity, ethnicity, sexual orientation, age, or disability, although they may express preference for projects proposed by residents of the United States. No person may be qualified as a Qualified Researcher if they act as an Agent of a foreign power ...*

The legislation includes a process of review and objection, with appeal processes, for platforms and researchers to agree on whether publication of research raises any of the kinds of issues we discuss above in this report. Notably, these are not restrictions solely to privacy laws (s 6(g)):

> *Qualified Platforms may object to the publication or release of any analysis that will necessarily expose Personal Information or otherwise violate federal, state, and local information sharing and privacy laws and regulations or any applicable rules, standards, regulations, and orders issued by the FTC. …*

Researchers who intentionally violate the act can be subject to civil and criminal enforcement under federal, state and local laws. If a qualified platform violates the legislation, then the FTC can take civil action imposing a civil penalty for each violation.

Conclusion on second class of issues

In this part we have explained the matters of law and public policy that we see presenting barriers to the adoption of enhanced access and transparency approaches. Importantly, however, we conclude these barriers exist whether the approach is collaborative and voluntary, or whether it is mandated through a legal framework. We think that policy makers or bodies such as GPAI can gain crucial insights from understanding the legal and practical impediments faced during previous collaborative transparency arrangements.

The barriers to enhanced access and transparency arrangements are both practical and principled. The strongest candidates for transparency laws still defer the resolution of many of these trade-offs to subsequent consultative processes and delegated legislation. The key insight is that transparency legislation may not be so easy to adopt and there is more conceptual work to be done. Advocates for enhanced transparency arrangements can enhance the likelihood that legislative measures will be

adopted successfully by doing this work, anticipating these barriers and doing what they can to work around them. GPAI's proposed methodology is one means of proactive research design in order to achieve this.

CONCLUSION

In this report, we have attempted to constructively contribute to GPAI's proposed approach by sharing insights on the legal and public policy issues raised by the kinds of projects described in the GPAI technical report. We emphasise that the two reports were written concurrently with limited opportunity for interaction between them. Again, it may be that the kinds of issues we raise here can be – or already have been – resolved by the specific research method that GPAI is proposing to adopt. Necessarily, we will be limited in our ability to assess the technical aspects of GPAI's proposal. Our focus sits with the legal and policy issues we can reasonably foresee arising. Importantly, whether or not those issues arise and whether they can be dealt with in practice rest heavily on the specifics of the particular proposal.

Notably, the issues we have covered sit in two classes. The first relate to the specific issues created by focusing on TVEC as the subject of research. The second class of issues relate to the legal and practical issues we foresee arising more generally in any attempt to pursue enhanced access and transparency to social media systems.

Nothing in the report above should be taken as an argument that social media recommender systems do or do not cause harm: it has been written specifically to avoid engaging with these questions. Instead, our focus has been on considering what kinds of issues might arise if an external party – whether researcher or regulator – were to seek to empirically investigate hypotheses about the impacts of social media companies' systems and their contribution to such harms. We strongly support such empirical investigations. Consistent with our support for those investigations, we have considered as far as we can in the context of this work what might be required to enable them to proceed. Any criticism of GPAI's approach must be assessed in that light.

By way of conclusion, we note that legislative approaches to transparency around the systems and impacts of private entities are deployed in a range of comparable policy areas. There are other areas of law and policy where: [82]

- States can show a demonstrable connection between a particular kind of conduct and a particular harm to a legitimate human rights interest.
- States use transparency and reporting based approaches to protect human rights or legitimate regulatory interests and to enable enforcement action.
- States are authorised to compel access to private information held by companies within the boundaries of acceptable access controls.

What is essential is that we remain cautious about the way that such regulatory approaches may impose unjustified limitations on existing rights and obligations, or may have second order effects we have not anticipated. From our investigations into this area, we have found that a human rights approach grounded in human rights instruments is a useful way of anchoring these issues to an existing framework and a common starting point for discussion among groups with potentially diverging priorities and expectations.

[82] We recommend MacCarthy M, 'Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry' (Transatlantic Working Group 2020).

ABOUT BRAINBOX

Brainbox is an independent consultancy and think tank based in New Zealand, which specialises in issues at the intersection of technology, politics, law and policy. Brainbox and its key personnel have prepared funded legal research reports and advice on the following subjects:

- A report to an investor coalition engaging with social media companies in response to the Christchurch Terror Attacks, assessing content moderation during objectionable content crises by Facebook, Alphabet and Twitter and the overall trajectory of regulation.
- The implementation of the law in digital systems and the representation and implementation of legal instruments in machine executable languages (Legislation as Code).
- The use of algorithmic methods to analyse written decisions by judicial bodies and the policy implications of this, including methods of enhancing access to primary legal materials (Judgments as Data).
- The relationship between concepts of "trust" and "automated decision-making", to support a wider research programme by the Digital Council for Aotearoa (Trust and Automated Decision Making).
- The legal implications of emerging technologies that create highly convincing but unreliable audio-visual media and how the New Zealand legal system deals with potentially harmful audio-visual content (Deepfakes and synthetic media).
- The policy implications of misinformation and disinformation, including attempts to regulate the creation and distribution of such information.
- A range of research investigations into health and disability policy, including how human rights instruments do or do not influence such policy (reports on accessibility, access to justice and human rights).

BIBLIOGRAPHY

'5Rights | The Digital Services Act Must Deliver for Children' <https://5rightsfoundation.com/in-action/the-digital-services-act-must-deliver-for-children.html> accessed 8 September 2021

'6.2m Tweets on EU Elections as Voters Turn to Twitter for Conversation' <https://blog.twitter.com/en_us/topics/company/2019/voters_turn_to_twitter_for_eu_elections.html> accessed 6 April 2021

'18 Trends That Highlight Fundamental Shifts in Culture' <https://blog.twitter.com/en_us/topics/insights/2019/18-trends-that-highlight-fundamental-shifts-in-culture.html> accessed 5 April 2021

'2019 El Paso Shooting', *Wikipedia* (2021) <https://en.wikipedia.org/w/index.php?title=2019_El_Paso_shooting&oldid=1015907913> accessed 6 April 2021

'2021 Global Summit' (*GIFCT*) <https://gifct.org/2021-global-summit/> accessed 6 September 2021

'A Further Update on New Zealand Terrorist Attack' (*About Facebook*, 21 March 2019) <https://about.fb.com/news/2019/03/technical-update-on-new-zealand/> accessed 15 March 2021

'A Look at the Research Behind Twitter Engage' <https://blog.twitter.com/en_us/a/2016/a-look-at-the-research-behind-twitter-engage.html> accessed 6 April 2021

'A Safer Internet for Europe, the Middle East and Africa' (*Google*, 11 February 2020) <https://blog.google/technology/safety-security/safer-internet-europe-middle-east-and-africa/> accessed 1 April 2021

'A Timeline of Recent Terrorist Attacks in Europe' (*Time*) <https://time.com/4607481/europe-terrorism-timeline-berlin-paris-nice-brussels/> accessed 1 April 2021

Abilov A and others, 'VoterFraud2020: A Multi-Modal Dataset of Election Fraud Claims on Twitter' [2021] arXiv:2101.08210 [cs] <http://arxiv.org/abs/2101.08210> accessed 6 April 2021

'About Tech Against Terrorism - Tech Against Terrorism' (4 September 2017) <https://www.techagainstterrorism.org/about/, https://www.techagainstterrorism.org/about/> accessed 18 March 2021

'Addressing Creator Feedback and an Update on My 2019 Priorities' (*blog.youtube*) <https://blog.youtube/inside-youtube/addressing-creator-feedback-and-update/> accessed 5 April 2021

'Addressing the Abuse of Tech to Spread Terrorist and Extremist Content' <https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c.html> accessed 6 April 2021

Alparslan Y and others, 'Towards Evaluating Gaussian Blurring in Perceptual Hashing as a Facial Image Filter' [2020] arXiv:2002.00140 [cs] <http://arxiv.org/abs/2002.00140> accessed 6 April 2021

Amarasingam DA, 'Turning the Tap Off: The Impacts of Social Media Shutdown After Sri Lanka's Easter Attacks' (*GNET*) <https://gnet-research.org/2021/03/05/turning-the-tap-off-the-impacts-of-social-media-shutdown-after-sri-lankas-easter-attacks/> accessed 13 April 2021

'An Open, Interconnected and Interoperable Internet (Joint Letter) | Global Partners Digital' <https://www.gp-digital.org/an-open-interconnected-and-interoperable-internet-joint-letter/> accessed 15 September 2021

‘An Update on Combating Hate and Dangerous Organizations’ (*About Facebook*, 12 May 2020) <https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/> accessed 9 March 2021

‘An Update on Our Commitment to Fight Terror Content Online’ (*blog.youtube*) <https://blog.youtube/news-and-events/an-update-on-our-commitment-to-fight-terror/> accessed 31 March 2021

‘An Update on Our Efforts to Combat Terrorism Online’ (*About Facebook*, 20 December 2019) <https://about.fb.com/news/2019/12/counterterrorism-efforts-update/> accessed 25 March 2021

‘An Update on Our Efforts to Combat Violent Extremism’ <https://blog.twitter.com/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html> accessed 6 April 2021

‘Announcing Google.Org’s New Safety Grants in Europe’ (*Google*, 4 February 2020) <https://blog.google/outreach-initiatives/google-org/announcing-googleorgs-new-safety-grants-europe/> accessed 1 April 2021

‘Arrested Coast Guard Officer Allegedly Planned Attack “On A Scale Rarely Seen”’ (*NPR.org*) <https://www.npr.org/2019/02/20/696470366/arrested-coast-guard-officer-planned-mass-terrorist-attack-on-a-scale-rarely-see> accessed 11 March 2021

‘At a Glance: Does the EU Digital Services Act Protect Freedom of Expression?’ (*ARTICLE 19*) <https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/> accessed 8 September 2021

Awan I, ‘Cyber-Extremism: Isis and the Power of Social Media’ (2017) 54 Society 138

Azarafrooz A and Brock J, ‘Fuzzy Hashing as Perturbation-Consistent Adversarial Kernel Embedding’ [2018] arXiv:1812.07071 [cs, stat] <http://arxiv.org/abs/1812.07071> accessed 6 April 2021

Baldi M and others, ‘On Fuzzy Syndrome Hashing with LDPC Coding’ [2011] arXiv:1107.1600 [cs, math] <http://arxiv.org/abs/1107.1600> accessed 6 April 2021

Banchik AV, ‘Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-Risk Human Rights–Related Content’ [2020] New Media & Society 1461444820912724

Basra R, ‘The YouTube Browsing Habits of a Lone-Actor Terrorist’ (*GNET*) <https://gnet-research.org/2020/06/22/the-youtube-browsing-habits-of-a-lone-actor-terrorist/> accessed 13 April 2021

‘Bear Witness, Take Action’ (*blog.youtube*) <https://blog.youtube/news-and-events/bear-witness-take-action/> accessed 1 April 2021

‘Behind the European Union’s Plan to Rewrite the Rules of Online Life’ (*Atlantic Council*, 26 June 2021) <https://www.atlanticcouncil.org/blogs/new-atlanticist/behind-the-european-unions-plan-to-rewrite-the-rules-of-online-life/> accessed 8 September 2021

‘Belarus Plane: What We Know and What We Don’t’ *BBC News* (25 June 2021) <https://www.bbc.com/news/world-europe-57239521> accessed 7 September 2021

Bellasio J and others, ‘Counterterrorism Evaluation: Taking Stock and Looking Ahead’ (RAND Corporation 2018) <https://www.rand.org/pubs/research_reports/RR2628.html> accessed 27 September 2021

Belli L, ‘Glossary on Platform Law and Policy Terms’ (*Internet Governance Forum*, 20 November 2020) <http://www.intgovforum.org/multilingual/content/glossary-on-platform-law-and-policy-terms> accessed 18 September 2021

Benesch S, ‘But Facebook’s Not a Country: How to Interpret Human Rights Law for Social Media Companies’ (2020) 38 Yale Journal on Regulation Bulletin 86

Bernstein J, 'Bad News: Selling the Story of Disinformation' (2021) September 2021 *Harper's Magazine* <https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/> accessed 6 September 2021

'Beyond Engagement: Aligning Algorithmic Recommendations With Prosocial Goals' (*Partnership on AI*, 21 January 2021) <https://partnershiponai.org/beyond-engagement-aligning-algorithmic-recommendations-with-prosocial-goals/> accessed 5 October 2021

Bhaskara VS and Bhattacharyya D, 'Emulating Malware Authors for Proactive Protection Using GANs over a Distributed Image Visualization of Dynamic File Behavior' [2018] arXiv:1807.07525 [cs, stat] <http://arxiv.org/abs/1807.07525> accessed 6 April 2021

'Big Tent Sendai: Smarter Ways to Share Information in a Crisis' (*Google*, 3 July 2012) <https://blog.google/outreach-initiatives/google-org/big-tent-sendai-smarter-ways-to-share/> accessed 1 April 2021

Bishop DP, 'Online Terrorist Content: Is It Time for an Independent Regulator?' (*GNET*) <https://gnet-research.org/2020/11/16/online-terrorist-content-is-it-time-for-an-independent-regulator/> accessed 13 April 2021

Biswas R and others, 'Perceptual Hashing Applied to Tor Domains Recognition' [2020] arXiv:2005.10090 [cs] <http://arxiv.org/abs/2005.10090> accessed 6 April 2021

Blake S, 'Embedded Blockchains: A Synthesis of Blockchains, Spread Spectrum Watermarking, Perceptual Hashing & Digital Signatures' [2020] arXiv:2009.00951 [cs, math] <http://arxiv.org/abs/2009.00951> accessed 6 April 2021

'Bot or Not? The Facts about Platform Manipulation on Twitter' <https://blog.twitter.com/en_us/topics/company/2020/bot-or-not.html> accessed 6 April 2021

Brad Smith, 'Microsoft, Other Tech Industry Leaders Team up with an International Coalition of Governments for a Multi-Stakeholder Solution' (*Microsoft On the Issues*, 24 September 2019) <https://blogs.microsoft.com/on-the-issues/2019/09/23/microsoft-other-tech-industry-leaders-team-up-with-an-international-coalition-of-governments-for-a-multi-stakeholder-solution/> accessed 8 March 2021

Bradley B, 'Doing Away with Harm' (2012) 85 Philosophy and phenomenological research 390

'Bringing Facebook Live to Android and More Countries' (*About Facebook*, 26 February 2016) <https://about.fb.com/news/2016/02/bringing-facebook-live-to-android-and-more-countries/> accessed 24 March 2021

'Bringing New Redirect Method Features to YouTube' (*blog.youtube*) <https://blog.youtube/news-and-events/bringing-new-redirect-method-features/> accessed 31 March 2021

'Bringing the Campaigning Power of Twitter and Counter-Narratives to Spain' <https://blog.twitter.com/en_us/a/2016/bringing-the-campaigning-power-of-twitter-and-counter-narratives-to-spain.html> accessed 5 April 2021

'Building a Safer Internet - Google Safety Center' <https://safety.google/engineering-center/> accessed 1 April 2021

'Building a Safer Internet, from Europe to Africa' (*Google*, 9 February 2021) <https://blog.google/technology/safety-security/building-safer-internet-europe-africa/> accessed 1 April 2021

Burke R, Felfernig A and Göker MH, 'Recommender Systems: An Overview' (2011) 32 AI Magazine 13

'Call for Comment on GDPR Article 40 Working Group' (*EDMO*) <https://edmo.eu/2020/11/24/call-for-comment-on-gdpr-article-40-working-group/> accessed 7 September 2021

‘Cambridge Analytica, GDPR - 1 Year on - a Lot of Words and Some Action’ (*Privacy International*) <http://privacyinternational.org/news-analysis/2857/cambridge-analytica-gdpr-1-year-lot-words-and-some-action> accessed 6 September 2021

Cambridge Consultants, on behalf of Ofcom UK, ‘Use of AI in Online Content Moderation’ (2019) <https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf> accessed 7 September 2021

Celma O and Lamere P, ‘If You Like Radiohead, You Might Like This Article’ (2011) 32 AI Magazine 57

Chase PH, ‘The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem’ (2019)

Chen P, ‘AI for Memers Part I’ (*JSK Class of 2020*, 12 December 2019) <https://medium.com/jsk-class-of-2020/pamela-chen-ai-for-memers-4ee32c9e6ae4> accessed 8 September 2021

Chilson N, ‘Seeing (Platforms) Like a State: Digital Legibility and Lessons for Platform Governance’ 29 33

‘Christchurch Call Community Consultation: Final Report’ (2021) <https://www.christchurchcall.com/christchurch-call-community-consultation-report.pdf> accessed 14 April 2021

‘Christchurch Earthquake — One Year Later: Live Streaming the Memorial Service on YouTube’ (*blog.youtube*) <https://blog.youtube/news-and-events/christchurch-earthquake-one-year-later/> accessed 31 March 2021

‘Christchurch Mosque Attack Livestream : Featured Classification Decisions : OFLC’ <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/> accessed 18 March 2021

‘Civil Society Positions on Christchurch Call Pledge’ <https://www.eff.org/files/2019/05/16/community_input_on_christchurch_call.pdf> accessed 6 April 2021

Cobbe J and Singh J, ‘Regulating Recommending: Motivations, Considerations, and Principles’ (2019) 10 European Journal of Law and Technology <https://ejlt.org/index.php/ejlt/article/view/686> accessed 8 September 2021

‘Code of Practice on Disinformation | Shaping Europe’s Digital Future’ <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation> accessed 1 April 2021

Coll S, ‘Alex Jones, the First Amendment, and the Digital Public Square’ (*The New Yorker*) <https://www.newyorker.com/magazine/2018/08/20/alex-jones-the-first-amendment-and-the-digital-public-square> accessed 31 March 2021

‘Combating Hate and Extremism’ (*About Facebook*, 17 September 2019) <https://about.fb.com/news/2019/09/combating-hate-and-extremism/> accessed 25 March 2021

‘Combating Violent Extremism’ <https://blog.twitter.com/en_us/a/2016/combating-violent-extremism.html> accessed 6 April 2021

Comerford M, ‘Two Years On: Understanding the Resonance of the Christchurch Attack on Imageboard Sites’ (*GNET*) <https://gnet-research.org/2021/03/24/two-years-on-understanding-the-resonance-of-the-christchurch-attack-on-imageboard-sites/> accessed 8 April 2021

‘Community Standards’ <https://m.facebook.com/communitystandards/additional_information/> accessed 1 April 2021

Conway M and others, 'Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts' (2019) 42 Studies in Conflict & Terrorism 141

Covington P, Adams J and Sargin E, 'Deep Neural Networks for YouTube Recommendations', *Proceedings of the 10th ACM Conference on Recommender Systems* (ACM 2016) <https://dl.acm.org/doi/10.1145/2959100.2959190> accessed 6 April 2021

'Creating a Dataset and a Challenge for Deepfakes' <https://ai.facebook.com/blog/deepfake-detection-challenge/> accessed 25 March 2021

Crenshaw M, *Countering Terrorism* (Brookings Institution Press 2017)

'Crisis Response' (*GIFCT*) <https://gifct.org/crisis-communications/> accessed 5 April 2021

'Csf-Final-Deck_03.26.19.Pdf' <https://about.fb.com/wp-content/uploads/2018/11/csf-final-deck_03.26.19.pdf> accessed 1 April 2021

Dalins J, Wilson C and Boudry D, 'PDQ & TMK + PDQF -- A Test Drive of Facebook's Perceptual Hashing Algorithms' [2019] arXiv:1912.07745 [cs] <http://arxiv.org/abs/1912.07745> accessed 6 April 2021

D'Anastasio C, 'A Christchurch Report Points to YouTube's Radicalization Trap' *Wired* <https://www.wired.com/story/christchurch-shooter-youtube-radicalization-extremism/> accessed 13 April 2021

Daphne Keller, 'Amplification and Its Discontents' (*Knight First Amendment Institute*, 8 June 2021) <https://knightcolumbia.org/content/amplification-and-its-discontents> accessed 7 September 2021

David Kaye, Joseph Cannataci, Fionnuala Ni Aolain, 'Mandates of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression; the Special Rapporteur on the Right to Privacy and the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism. Letter Regarding the European Commission's Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online to Complement Directive 2017/541 on Combating Terrorism. OL OTH 71/2018' (7 December 2018) <https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234> accessed 6 September 2021

Davidson J and others, 'The YouTube Video Recommendation System' (2010)

De Herve JDG and others, 'A Perceptual Hash Function to Store and Retrieve Large Scale DNA Sequences' [2014] arXiv:1412.5517 [cs, q-bio] <http://arxiv.org/abs/1412.5517> accessed 6 April 2021

'December 2020 Coordinated Inauthentic Behavior Report' (*About Facebook*, 12 January 2021) <https://about.fb.com/news/2021/01/december-2020-coordinated-inauthentic-behavior-report/> accessed 25 March 2021

'Defending the Truth of the Holocaust in 2021' (*blog.youtube*) <https://blog.youtube/news-and-events/defending-the-truth-holocaust-2021/> accessed 31 March 2021

'Designing with Constraint: Twitter's Approach to Email' <https://blog.twitter.com/en_us/a/2015/designing-with-constraint-twitters-approach-to-email.html> accessed 5 April 2021

'Despite A Ban, Facebook Continued To Label People As Interested In Militias For Advertisers' (*BuzzFeed News*) <https://www.buzzfeednews.com/article/ryanmac/facebook-militia-interest-category-advertisers-ban> accessed 9 April 2021

'Dialogue with Sen. Lieberman on Terrorism Videos' (*blog.youtube*) <https://blog.youtube/news-and-events/dialogue-with-sen-lieberman-on/> accessed 31 March 2021

‘Digital Services Act: Civil Liberties Committee Pushes for Digital Privacy and Free Speech | European Pirate Party’ <https://european-pirateparty.eu/dsa-pirate-position-adopted/> accessed 8 September 2021

‘Discord Chats May Be Crucial to Lawsuits over Neo-Nazi Violence’ (*Engadget*) <https://www.engadget.com/2017-08-26-discord-chats-may-help-charlottesville-lawsuits.html> accessed 11 March 2021

Douek E, ‘Facebook’s “Oversight Board:” Move Fast With Stable Infrastructure And Humility’ 21 N.C. J. L. & Tech. 1 (2019)

——, ‘Australia’s “Abhorrent Violent Material” Law: Shouting “Nerd Harder” and Drowning Out Speech’ (Social Science Research Network 2019) SSRN Scholarly Paper ID 3443220 <https://papers.ssrn.com/abstract=3443220> accessed 6 April 2021

——, ‘The Free Speech Blind Spot: Foreign Election Interference On Social Media’ [2020] DRAFT – COMBATING ELECTION INTERFERENCE WHEN FOREIGN POWERS TARGET DEMOCRACIES (Duncan B. Hollis & Jens David Ohlin eds., Oxford University Press, forthcoming 2020) 27

——, ‘The Limits of International Law in Content Moderation’ [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3709566> accessed 31 March 2021

——, ‘The Rise of Content Cartels’ [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3572309> accessed 31 March 2021

——, ‘Governing Online Speech: From “Posts-As-Trumps” To Proportionality & Probability’ (2021) 121(3) Columbia Law Review 759

‘Draft Online Safety Bill’ (*GOV.UK*) <https://www.gov.uk/government/publications/draft-online-safety-bill> accessed 16 September 2021

Echikson W and Knodt O, ‘Germany’s NetzDG: A Key Test for Combatting Online Hate’ (Social Science Research Network 2018) SSRN Scholarly Paper ID 3300636 <https://papers.ssrn.com/abstract=3300636> accessed 6 April 2021

Edelman G, ‘What Social Media Needs to Learn From Traditional Media’ *Wired* <https://www.wired.com/story/what-social-media-needs-to-learn-from-traditional-media/> accessed 23 September 2021

Elhai W, ‘Regulating Digital Harm Across Borders: Exploring a Content Platform Commission’, *International Conference on Social Media and Society* (Association for Computing Machinery 2020) <https://doi.org/10.1145/3400806.3400832> accessed 5 April 2021

‘Empowering Dutch NGOs to Amplify Their Voice on Twitter’ <https://blog.twitter.com/en_us/a/2016/empowering-dutch-ngos-to-amplify-their-voice-on-twitter.html> accessed 5 April 2021

Erin Saltman, ‘Countering Terrorism and Violent Extremism at Facebook: Technology, Expertise and Partnerships’ in Maya Mirchandani (ed), *Tackling Insurgent Ideologies in a Pandemic World* (2020)

‘EU Counterterrorism Directive Seriously Flawed’ (*Human Rights Watch*, 30 November 2016) <https://www.hrw.org/news/2016/11/30/eu-counterterrorism-directive-seriously-flawed> accessed 6 September 2021

‘EU: Regulation of Recommender Systems in the Digital Services Act’ (*ARTICLE 19*) <https://www.article19.org/resources/eu-regulation-of-recommender-systems-in-the-digital-services-act/> accessed 8 September 2021

European Commission, ‘Proposal for a Regulation Of The European Parliament And Of The Council on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC (Explanatory Memorandum)’ (15 December 2020) <https://eur-

lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en> accessed 7 September 2021

'European Council Conclusions on Security and Defence, 22/06/2017' <https://www.consilium.europa.eu/en/press/press-releases/2017/06/22/euco-security-defence/> accessed 1 April 2021

'Evaluating Platform Accountability: Terrorist Content on YouTube - Dhiraj Murthy, 2021' <https://journals-sagepub-com.ezproxy.otago.ac.nz/doi/10.1177/0002764221989774> accessed 6 April 2021

'Explainers' (*GNET*) <https://gnet-research.org/explainers/> accessed 18 March 2021

'Exposed Email Logs Show 8kun Owner in Contact With QAnon Influencers and Enthusiasts' (*bellingcat*, 7 January 2021) <https://www.bellingcat.com/news/2021/01/07/exposed-email-logs-show-8kun-owner-in-contact-with-qanon-influencers-and-enthusiasts/> accessed 11 March 2021

'Extracts From ISD's Submitted Response to the UK Government Online Harms White Paper' (*ISD*) <https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/> accessed 8 September 2021

Facebook and others, 'One Dead, Three Injured in Poway Synagogue Shooting' (*San Diego Union-Tribune*, 27 April 2019) <https://www.sandiegouniontribune.com/news/public-safety/story/2019-04-27/reports-of-several-people-shot-at-poway-synagogue> accessed 11 March 2021

'Facebook at UNGA 2020' (*About Facebook*, 21 September 2020) <https://about.fb.com/news/2020/09/facebook-at-unga-2020/> accessed 25 March 2021

'Facebook Executive In 2016: "Maybe Someone Dies In A Terrorist Attack Coordinated On Our Tools"' (*BuzzFeed News*) <https://www.buzzfeednews.com/article/ryanmac/growth-at-any-cost-top-facebook-executive-defended-data> accessed 23 September 2021

'Facebook Joins Other Tech Companies to Support the Christchurch Call to Action' (*About Facebook*, 15 May 2019) <https://about.fb.com/news/2019/05/christchurch-call-to-action/> accessed 9 March 2021

'Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism' (*About Facebook*, 26 June 2017) <https://about.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/> accessed 1 April 2021

'——' (*blog.youtube*) <https://blog.youtube/news-and-events/facebook-microsoft-twitter-and-youtube/> accessed 1 April 2021

'Facebook Rolls Out News Feed Change That Blocks Watchdogs from Gathering Data – The Markup' <https://themarkup.org/citizen-browser/2021/09/21/facebook-rolls-out-news-feed-change-that-blocks-watchdogs-from-gathering-data> accessed 6 October 2021

'Facebook Says New Rule Would Have Stopped Christchurch Shooter Livestreaming' (*Stuff*, 15 May 2019) <https://www.stuff.co.nz/national/politics/112756865/facebook-says-new-rule-would-have-stopped-christchurch-shooter-livestreaming> accessed 14 April 2021

'Facebook Says No One Flagged NZ Mosque Shooting Livestream' (*The Salt Lake Tribune*) <https://sltrib.com/news/nation-world/2019/03/19/facebook-says-no-one> accessed 18 March 2021

'Facebook Whistleblower Says Company Incentivizes "Angry, Polarizing, Divisive Content" - 60 Minutes - CBS News' <https://www.cbsnews.com/news/facebook-whistleblower-polarizing-divisive-content-60-minutes-2021-10-03/> accessed 4 October 2021

'Facebook's Community Standards: How and Where We Draw the Line' (*About Facebook*, 23 May 2017) <https://about.fb.com/news/2017/05/facebooks-community-standards-how-and-where-we-draw-the-line/> accessed 24 March 2021

‘First Grants Announced for Independent Research on Social Media’s Impact on Democracy Using Facebook Data’ <https://socialscience.one/blog/first-grants-announced-independent-research-social-media%E2%80%99s-impact-democracy> accessed 8 September 2021

‘Five Tips for Brands from #Twitter4Politics’ <https://blog.twitter.com/en_us/a/2015/five-tips-for-brands-from-twitter4politics.html> accessed 6 April 2021

Ford P, ‘Combatting Terrorist Propaganda’ (2020) 15 Journal of Policing, Intelligence and Counter Terrorism 175

‘Four Steps We’re Taking Today to Fight Terrorism Online’ (*Google*, 18 June 2017) <https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/> accessed 31 March 2021

‘Fourth Intermediate Results of the EU Code of Practice against Disinformation | Shaping Europe’s Digital Future’ <https://wayback.archive-it.org/12090/20210620011601/https://digital-strategy.ec.europa.eu/en/news/fourth-intermediate-results-eu-code-practice-against-disinformation> accessed 8 September 2021

‘Frequently Asked Questions (FAQ) - Tech Against Terrorism’ (28 November 2017) <https://www.techagainstterrorism.org/about/faq/, https://www.techagainstterrorism.org/about/faq/> accessed 18 March 2021

‘From Countering Radicalization to Disrupting Illicit Networks: What’s next for Google Ideas’ (*Official Google Blog*) <https://googleblog.blogspot.com/2012/04/from-countering-radicalization-to.html> accessed 1 April 2021

Fukuyama F, ‘The Future of Platform Power: Solving for a Moving Target’ (2021) 32 Journal of Democracy 173

‘Funders Are Ready To Pull Out Of Facebook’s Academic Data Sharing Project’ (*BuzzFeed News*) <https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data> accessed 8 September 2021

Ganesh B and Bright J, ‘Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation’ (2020) 12 Policy & Internet 6

——, *Extreme Digital Speech: Contexts, Responses, and Solutions* (VOX-Pol Network of Excellence 2020)

‘Getting Input on an Oversight Board’ (*About Facebook*, 1 April 2019) <https://about.fb.com/news/2019/04/input-on-an-oversight-board/> accessed 5 April 2021

Ghosh D and Srinivasan R, ‘The Future of Platform Power: Reining In Big Tech’ (2021) 32 Journal of Democracy 163

GIFCT, ‘Transparency Recommendations for GIFCT: Prepared by the GIFCT Transparency Working Group’ (2021) <https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyRecommendations.pdf> accessed 7 September 2021

——, ‘Content-Sharing Algorithms, Processes, and Positive Interventions Working Group. Part 1: Content-Sharing Algorithms & Processes’ (2021) <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI1-2021.pdf> accessed 7 September 2021

——, ‘Content-Sharing Algorithms, Processes, and Positive Interventions Working Group. Part 2: Positive Interventions’ (2021) <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI2-2021.pdf> accessed 7 September 2021

‘GIFCT Transparency Report, July 2020’ <https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf> accessed 5 April 2021

‘GIFCT Transparency Working Group: One-Year Review of Discussions’ (2021) <https://gifct.org/wp-content/uploads/2021/07/GIFCT-WorkingGroup21-OneYearReview.pdf> accessed 7 September 2021

Gillespie T, ‘The Politics of “Platforms”’ (2010) 12 New media & society 347

——, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Kindle edition, Yale University Press 2018)

‘Global Internet Forum to Counter Terrorism: An Update on Our Progress’ (*blog.youtube*) <https://blog.youtube/news-and-events/global-internet-forum-to-counter/> accessed 1 April 2021

‘Global Internet Forum To Counter Terrorism: An Update on Our Progress Two Years On’ (*About Facebook*, 25 July 2019) <https://about.fb.com/news/2019/07/global-internet-forum-to-counter-terrorism-an-update-on-our-progress-two-years-on/> accessed 25 March 2021

‘Global Internet Forum to Counter Terrorism to Hold First Meeting’ <https://blog.twitter.com/en_us/topics/insights/2017/Global-Internet-Forum-to-Counter-Terrorism-to-hold-first-meeting.html> accessed 6 April 2021

‘Global Internet Forum to Counter Terrorism to Hold First Meeting in San Francisco’ (*About Facebook*, 31 July 2017) <https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco/> accessed 25 March 2021

‘——’ (*About Facebook*, 31 July 2017) <https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco/> accessed 5 April 2021

‘Global Internet Forum to Counter Terrorism to Hold First Meeting in San Francisco’ (*blog.youtube*) <https://blog.youtube/news-and-events/global-internet-forum-san-francisco/> accessed 1 April 2021

Global Network Initiative, ‘Content Regulation and Human Rights: Analysis and Recommendations’ (October 2020) <https://globalnetworkinitiative.org/wp-content/uploads/2020/10/GNI-Content-Regulation-HR-Policy-Brief.pdf> accessed 20 September 2021

‘Global Network Initiative Releases 2015 Assessment Report’ (*Google*, 8 July 2016) <https://blog.google/outreach-initiatives/public-policy/global-network-initiative-releases-2015/> accessed 1 April 2021

‘GNI Resignation Letter’ (*Electronic Frontier Foundation*, 9 October 2013) <https://www.eff.org/document/gni-resignation-letter> accessed 6 April 2021

‘Google Ideas: Joining the Fight against Drug Cartels and Other Illicit Networks’ (*Google*, 16 July 2012) <https://blog.google/alphabet/google-ideas-joining-fight-against-drug/> accessed 1 April 2021

‘Google Ideas Launches Summit Against Violent Extremism’ (*Google Europe Blog*) <https://europe.googleblog.com/2011/06/google-ideas-launches-summit-against.html> accessed 1 April 2021

‘Google-Funded Report on White Supremacy Downplays YouTube’s Role in Driving People to Extremism’ (*Stuff*, 16 December 2020) <https://www.stuff.co.nz/technology/300185901/googlefunded-report-on-white-supremacy-downplays-youtubes-role-in-driving-people-to-extremism> accessed 5 April 2021

‘Google.Org Impact Challenge on Safety’ (*Google.org Impact Challenge on Safety*) <https://impactchallenge.withgoogle.com/safety2019> accessed 1 April 2021

Gorwa R, ‘What Is Platform Governance?’ (2019) 22 Information, Communication & Society 854

——, ‘The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content’ (2019) 8 Internet Policy Review <https://policyreview.info/node/1407> accessed 6 April 2021

Gorwa R, Binns R and Katzenbach C, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945

Graham Smith (Cyberleagle), 'An Online Harms Compendium' <https://www.cyberleagle.com/2020/02/an-online-harms-compendium.html> accessed 16 September 2021

Greenfield S, 'Social Media Platforms: Preserving Evidence of International Crimes Notes' (2018) 2 International Comparative, Policy & Ethics Law Review 821

Grimmelmann J, 'The Virtues of Moderation' (LawArXiv 2017) preprint <https://osf.io/qwxf5> accessed 16 March 2021

'GSEC Dublin: A Content Responsibility Center for Europe' (*Google*, 27 January 2021) <https://blog.google/around-the-globe/google-europe/gsec-dublin-content-responsibility-center-europe/> accessed 1 April 2021

'Guest Post: Why We Must Remember the Holocaust' <https://blog.twitter.com/en_us/topics/events/2019/we_remember.html> accessed 6 April 2021

Hans Bredow Institute, 'Setting Rules for 2.7 Billion: A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study' (January 2020) <https://www.hans-bredow-institut.de/uploads/media/Publikationen/cms/media/7mkl6yl_AP_WiP001InsideFacebook.pdf> accessed 1 April 2021

'Hard Questions: How We Counter Terrorism' (*About Facebook*, 15 June 2017) <https://about.fb.com/news/2017/06/how-we-counter-terrorism/> accessed 25 March 2021

'Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok' (*ISD*) <https://www.isdglobal.org/isd-publications/hatescape-an-in-depth-analysis-of-extremism-and-hate-speech-on-tiktok/> accessed 18 September 2021

Heidi Tworek, 'Social Media Councils' (*Centre for International Governance Innovation*, 28 October 2019) <https://www.cigionline.org/articles/social-media-councils> accessed 14 April 2021

Heller B, 'Combating Terrorist-Related Content Through AI and Information Sharing' (The Carr Center for Human Rights Policy, Harvard University 2019)

Hendrix J, 'A Whistleblower, Facebook, Social Media & Polarization' (*Tech Policy Press*, 4 October 2021) <https://techpolicy.press/a-whistleblower-facebook-social-media-polarization/> accessed 5 October 2021

Horwitz J, 'Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.' *Wall Street Journal* (13 September 2021) <https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353> accessed 15 September 2021

'How Does Instagram Decide Which Ads to Show Me? | Instagram Help Centre' <https://help.instagram.com/173081309564229?helpref=related> accessed 6 September 2021

'How Facebook Got Addicted to Spreading Misinformation' (*MIT Technology Review*) <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/> accessed 6 September 2021

'How Facebook Relies on Accenture to Scrub Toxic Content - The New York Times' <https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html> accessed 6 September 2021

'How Twitter Is Fighting Spam and Malicious Automation' <https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html> accessed 6 April 2021

‘How We’re Supporting Smart Regulation and Policy Innovation in 2019’ (*Google*, 8 January 2019) <https://blog.google/perspectives/kent-walker/principles-evolving-technology-policy-2019/> accessed 5 April 2021

‘How YouTube Supports Elections’ (*blog.youtube*) <https://blog.youtube/news-and-events/how-youtube-supports-elections/> accessed 1 April 2021

‘Human Rights Impact Assessment: Global Internet Forum to Counter Terrorism | Reports | BSR’ <https://www.bsr.org/en/our-insights/report-view/human-rights-impact-assessment-global-internet-forum-to-counter-terrorism> accessed 6 September 2021

‘“I Have Blood On My Hands”: A Whistleblower Says Facebook Ignored Global Political Manipulation’ (*BuzzFeed News*) <https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo> accessed 6 September 2021

‘Inclusion & Diversity Report May 2019’ <https://blog.twitter.com/en_us/topics/company/2019/Board-Update-Inclusion-Diversity-Report-May2019.html> accessed 6 April 2021

‘#Influencer Voices: How the National Association of Manufacturers Captures Attention on Twitter during Major Political Events’ <https://blog.twitter.com/en_us/a/2015/influencer-voices-how-the-national-association-of-manufacturers-captures-attention-on-twitter.html> accessed 5 April 2021

Initiative GN, ‘The DSA: An Opportunity to Build Human Rights Safeguards into Notice and Action by Emma Llansó’ (*The GNI Blog*, 17 August 2020) <https://medium.com/global-network-initiative-collection/the-dsa-an-opportunity-to-build-human-rights-safeguards-into-notice-and-action-by-emma-llans%C3%B3-e0487397646f> accessed 8 September 2021

‘Insights from the 17th Twitter Transparency Report’ <https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html> accessed 6 April 2021

‘Instagram Keeps a Detailed List of Everything It Thinks You’re Interested in — Here’s How to Find It’ (*Business Insider Australia*, 11 June 2019) <https://www.businessinsider.com.au/instagram-interests-list-how-to-find-2019-6> accessed 6 September 2021

‘Introducing Event Targeting’ <https://blog.twitter.com/en_us/a/2015/introducing-event-targeting.html> accessed 5 April 2021

‘Introducing Live Video and Collages’ (*About Facebook*, 3 December 2015) <https://about.fb.com/news/2015/12/introducing-live-video-and-collages/> accessed 24 March 2021

‘Introducing the Developer Platform’s Academic Research Advisory Board’ <https://blog.twitter.com/developer/en_us/topics/community/2021/introducing-the-developer-platform-academic-research-advisory-board> accessed 2 September 2021

‘Introducing the New Twitter Transparency Center’ <https://blog.twitter.com/en_us/topics/company/2020/new-transparency-center.html> accessed 6 April 2021

‘Investigating Information Operations in West Papua: A Digital Forensic Case Study of Cross-Platform Network Analysis’ (*bellingcat*, 11 October 2019) <https://www.bellingcat.com/news/rest-of-world/2019/10/11/investigating-information-operations-in-west-papua-a-digital-forensic-case-study-of-cross-platform-network-analysis/> accessed 1 April 2021

Jackson S, ‘The Double-Edged Sword of Banning Extremists from Social Media’ (SocArXiv, 2 July 2019) <https://osf.io/preprints/socarxiv/2g7yd/> accessed 6 April 2021

Jie Z, ‘A Novel Block-DCT and PCA Based Image Perceptual Hashing Algorithm’ [2013] arXiv:1306.4079 [cs] <http://arxiv.org/abs/1306.4079> accessed 6 April 2021

‘Jigsaw’ (*Jigsaw*) <https://jigsaw.google.com/> accessed 1 April 2021

——, ‘Google’s Jigsaw Announces Toxicity-Reducing API, Perspective, Is Processing 500M Requests Daily’ <https://www.prnewswire.com/news-releases/googles-jigsaw-announces-toxicity-reducing-api-perspective-is-processing-500m-requests-daily-301223600.html> accessed 7 September 2021

‘Jimmy Wales on Systems and Incentives (Ep. 109)’ <https://conversationswithtyler.com/episodes/jimmy-wales/> accessed 24 March 2021

‘Joint Letter to New Executive Director, Global Internet Forum to Counter Terrorism’ (*Human Rights Watch*, 30 July 2020) <https://www.hrw.org/news/2020/07/30/joint-letter-new-executive-director-global-internet-forum-counter-terrorism> accessed 13 April 2021

Jørgensen RF, ‘What Platforms Mean When They Talk About Human Rights’ (2017) 9 Policy & Internet 280

@jshermcyber, ‘The Christchurch Report Points to Better Avenues for Internet Reform’ (*Lawfare*, 26 March 2021) <https://www.lawfareblog.com/christchurch-report-points-better-avenues-internet-reform> accessed 9 April 2021

Kate Klonick, ‘Facebook Released Its Content Moderation Rules. Now What?’ New York Times, 26 April 2018

Kate Klonick and Thomas Kadri, ‘How to Make Facebook’s “Supreme Court” Work’, New York Times, 17 November 2018

Katzenbach C and Ulbricht L, ‘Algorithmic Governance’ (2019) 8 Internet Policy Review <https://policyreview.info/node/1424> accessed 6 April 2021

Keen F, ‘Online Subcultures and the Challenges of Moderation’ (*GNET*) <https://gnet-research.org/2020/10/01/online-subcultures-and-the-challenges-of-moderation/> accessed 13 April 2021

‘Keeping Our Users Secure’ <https://blog.twitter.com/en_us/a/2013/keeping-our-users-secure.html> accessed 5 April 2021

Keller D, ‘Making Google the Censor’ (*New York Times (Online)*, 12 June 2017) <http://search.proquest.com/docview/1908292276/abstract/3629B3851B694C5FPQ/1> accessed 31 March 2021

——, ‘Don’t Force Google to Export Other Countries’ Laws’ (*New York Times (Online)*, 10 September 2018) <http://search.proquest.com/docview/2101581443/abstract/787930A6ED0444B4PQ/1> accessed 31 March 2021

——, ‘The Stubborn, Misguided Myth That Internet Platforms Must Be “Neutral”’ *The Washington post* (Washington, DC, 2019)

——, ‘Facebook Filters, Fundamental Rights, and the CJEU’s Glawischnig-Piesczek Ruling’ (2020) 69 GRUR International 616

——, ‘The Future of Platform Power: Making Middleware Work’ (2021) 32 Journal of Democracy 168

Keller D and Brown BD, ‘Europe’s Web Privacy Rules: Bad for Google, Bad for Everyone’ (*New York Times (Online)*, 25 April 2016) <http://search.proquest.com/docview/1783859103/abstract/F9541C172E5E42E0PQ/1> accessed 31 March 2021

Kitchin R, ‘Thinking Critically about and Researching Algorithms’ (2017) 20 Information, Communication & Society 14

Klonick K, ‘Re-Shaming the Debate: Social Norms, Shame and Regulation in an Internet Age’ 75 Md. L. Rev. 1029 (2016)

——, ‘Networked Technologies’ Transformation of Social Norms, Private Self-Regulation, and the Law’ (ProQuest Dissertations Publishing 2018) <https://search.proquest.com/docview/2088928846?pq-origsite=primo> accessed 6 April 2021

——, ‘The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression’ [2020] The Yale Law Journal 83

——, ‘Content Moderation Modulation: Deliberating on How to Regulate--or Not Regulate--Online Speech in the Era of Evolving Social Media’ (2021) 64 Communications of the ACM 29

——, ‘Inside the Making of Facebook’s Supreme Court’ (*The New Yorker*) <https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court> accessed 31 March 2021

——, ‘Inside the Team at Facebook That Dealt with the Christchurch Shooting’ (*The New Yorker*) <https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting> accessed 31 March 2021

——, ‘The New Governors: The People, Rules, And Processes Governing Online Speech’ 131 Harv. L. Rev. 1598

Klonick K and Kadri T, ‘How to Make Facebook’s “Supreme Court” Work’ (*New York Times (Online)*, 17 November 2018) <http://search.proquest.com/docview/2134340716/abstract/6A651F6F91E04E97PQ/1> accessed 31 March 2021

‘Korero Whakamauahara - Hate Speech (New Zealand Human Rights Commission, 2019)’ <https://www.hrc.co.nz/files/2915/7653/6167/Korero_Whakamauahara-_Hate_Speech_FINAL_13.12.2019.pdf> accessed 8 March 2021

Emily Kubin & Christian von Sikorski (2021) The role of (social) media in political polarization: a systematic review, Annals of the International Communication Association, DOI: 10.1080/23808985.2021.1976070

Langvardt K, ‘Regulating Online Content Moderation’ (2018) 106 The Georgetown law journal 1353

‘Lawmakers Want to Force Big Tech to Give Researchers More Data’ (*Protocol — The people, power and politics of tech*, 20 May 2021) <https://www.protocol.com/policy/social-media-data-act> accessed 18 August 2021

Leader Maynard J and Benesch S, ‘Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention’ (2016) 9 Genocide studies and prevention 70

Lee E, ‘Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance’ (2020) 70 American University Law Review 913

Leman J and Pektaş Ş, *Militant Jihadism: Today and Tomorrow* (Leuven University Press 2019)

Lewis R, ‘Broadcasting the Reactionary Right on YouTube’ (Data & Society Research Institute)

Li K, ‘Hashing for Multimedia Similarity Modeling and Large-Scale Retrieval’ (University of Central Florida 2017)

Li Y, Jang J and Ou X, ‘Topology-Aware Hashing for Effective Control Flow Graph Similarity Analysis’ [2020] arXiv:2004.06563 [cs] <http://arxiv.org/abs/2004.06563> accessed 6 April 2021

Llansó E and others, ‘Artificial Intelligence, Content Moderation and Freedom of Expression’ (Transatlantic Working Group 2020)

Lorna Woods and William Perrin, ‘Online Harm Reduction – a Statutory Duty of Care and Regulator’ (Carnegie UK Trust 2019)

https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf> accessed 8 September 2021

Lyons M, 'Regulating Terrorist Content Online: Considerations and Trade-Offs' (*Counter Terror Business*, 14 October 2019) <https://counterterrorbusiness.com/features/regulating-terrorist-content-online-considerations-and-trade-offs> accessed 6 April 2021

Mac R, 'Facebook Apologizes After A.I. Puts "Primates" Label on Video of Black Men' *The New York Times* (3 September 2021) <https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html> accessed 6 September 2021

MacCarthy M, 'Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry' (Transatlantic Working Group 2020)

'Machine Learning Can Identify Weapons in the Christchurch Attack Video' <https://www.vice.com/en/article/xwnzz4/machine-learning-artificial-intelligence-christchurch-attack-video-facebook-amazon-rekognition> accessed 18 March 2021

Mackey A, 'The PACT Act's Attempt to Help Internet Users Hold Platforms Accountable Will End Up Hurting Online Speakers' (*Electronic Frontier Foundation*, 21 July 2020) <https://www.eff.org/deeplinks/2020/07/pact-acts-attempt-help-internet-users-hold-platforms-accountable-will-end-hurting> accessed 6 October 2021

'Making It Easier to Report Threats to Law Enforcement' <https://blog.twitter.com/en_us/a/2015/making-it-easier-to-report-threats-to-law-enforcement.html> accessed 6 April 2021

'Making Our Rules Easier to Understand' <https://blog.twitter.com/en_us/topics/company/2019/rules-refresh.html> accessed 6 April 2021

'"Mark Changed The Rules": How Facebook Went Easy On Alex Jones And Other Right-Wing Figures' (*BuzzFeed News*) <https://www.buzzfeednews.com/article/ryanmac/mark-zuckerberg-joel-kaplan-facebook-alex-jones> accessed 7 September 2021

'Mark Warner Is Ready to Fight for Section 230 Reform' (*Protocol — The people, power and politics of tech*, 22 March 2021) <https://www.protocol.com/policy/mark-warner-section-230> accessed 6 April 2021

Mark Zuckerberg, 'Some Thoughts on Facebook and the Election' (13 November 2016) <https://www.facebook.com/zuck/posts/10103253901916271> accessed 24 March 2021

'Mark Zuckerberg Says Fake News On Facebook Didn't Change The Election' (*BuzzFeed News*) <https://www.buzzfeednews.com/article/stephaniemlee/zuckerberg-techonomy-fake-news-election> accessed 24 March 2021

Martin Scheinin, 'Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism, Martin Scheinin' <https://primarysources.brillonline.com/browse/human-rights-documents-online/promotion-and-protection-of-all-human-rights-civil-political-economic-social-and-cultural-rights-including-the-right-to-development;hrdhrd99702016149> accessed 6 September 2021

'Mass Violence, Extremism, and Digital Responsibility' (*U.S. Senate Committee on Commerce, Science, & Transportation*, 18 September 2019) <https://www.commerce.senate.gov/2019/9/mass-violence-extremism-and-digital-responsibility> accessed 6 April 2021

'Mastodon' <https://joinmastodon.org/> accessed 11 March 2021

Mattheis A, 'Beyond the "LULZ:" Memifying Murder as "Meaningful" Gamification in Far-Right Content' (*GNET*) <https://gnet-research.org/2021/01/18/beyond-the-lulz-memifying-murder-as-meaningful-gamification-in-far-right-content/> accessed 13 April 2021

McSherry JCY and C, ‘Content Moderation Is Broken. Let Us Count the Ways.’ (*Electronic Frontier Foundation*, 29 April 2019) <https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways> accessed 7 September 2021

‘Meet the Teams Keeping Our Corner of the Internet Safer’ (*Google*, 5 February 2019) <https://blog.google/around-the-globe/google-europe/meet-teams-keeping-our-corner-internet-safer/> accessed 1 April 2021

Messing S and others, ‘Facebook URL Shares’ <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EIAACS> accessed 8 September 2021

‘Microsoft Partners with Institute for Strategic Dialogue and NGOs to Discourage Online Radicalization to Violence’ (*Microsoft On the Issues*, 18 April 2017) <https://blogs.microsoft.com/on-the-issues/2017/04/18/microsoft-partners-institute-strategic-dialogue-ngos-discourage-online-radicalization-violence/> accessed 1 April 2021

Montgomery M, ‘Disinformation as a Wicked Problem: Why We Need Co-Regulatory Frameworks’ 14

Moscow MB, ‘Russian Software to Spot “Deviant” Thinking Misses Two Mass Shootings’ <https://www.thetimes.co.uk/article/russian-software-to-spot-deviant-thinking-misses-two-mass-shootings-58kp0rqb9> accessed 12 October 2021

Murthy D, ‘Evaluating Platform Accountability: Terrorist Content on YouTube’ [2021] American Behavioral Scientist 0002764221989774

‘/N/ - New Zealand Mobile Carriers Block 8chan, 4chan, and LiveLeak’ (18 March 2019) <https://web.archive.org/web/20190318033153/https:/8ch.net/n/res/756614.html> accessed 11 March 2021

‘New Progress in Using AI to Detect Harmful Content’ <https://ai.facebook.com/blog/community-standards-report/> accessed 13 April 2021

‘Next Steps for the Global Internet Forum to Counter Terrorism’ (*About Facebook*, 23 September 2019) <https://about.fb.com/news/2019/09/next-steps-for-gifct/> accessed 25 March 2021

Nguyen DT and others, ‘Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises’ [2017] arXiv:1704.02602 [cs] <http://arxiv.org/abs/1704.02602> accessed 6 April 2021

‘OHCHR | Report on Encryption, Anonymity, and the Human Rights Framework’ <https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/CallForSubmission.aspx> accessed 10 September 2021

‘On YouTube’s Recommendation System’ (*blog.youtube*) <https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/> accessed 18 September 2021

‘Op Ed: To Keep Social Media from Inciting Violence, Focus on Responses to Posts More than the Posts Themselves | Dangerous Speech Project’ (*Dangerous Speech Project* |, 31 May 2021) <https://dangerousspeech.org/to-keep-social-media-from-inciting-violence-focus-on-responses-to-posts-more-than-the-posts-themselves/> accessed 7 September 2021

‘Operating with Impunity - Hateful Extremism: The Need for a Legal Framework’ (Commission for Countering Extremism 2021)

‘Opinion | How to Make Facebook’s “Supreme Court” Work - The New York Times’ <https://www.nytimes.com/2018/11/17/opinion/facebook-supreme-court-speech.html> accessed 6 April 2021

OseiTutu JJ, ‘Corporate “Human Rights” to Intellectual Property Protection?’ (2015) 55 Santa Clara law review 1

Osnos E, 'How to Talk About the New Zealand Massacre: More Sunlight, Less Oxygen' (*The New Yorker*) <https://www.newyorker.com/news/daily-comment/how-to-talk-about-the-new-zealand-massacre-more-sunlight-less-oxygen> accessed 31 March 2021

'Our #DataForGood Partnership with New Zealand's NCPACS' <https://blog.twitter.com/en_us/topics/company/2020/christchurch-otago-nspacs.html> accessed 6 April 2021

'Our Ongoing Work to Tackle Hate' (*blog.youtube*) <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/> accessed 9 March 2021

'Oversight Board, Case Decision 2021-001-FB-FBR' (6 May 2021) <https://www.oversightboard.com/decision/FB-691QAMHJ> accessed 7 September 2021

'Oversight Board Overturns Original Facebook Decision: Case 2021-009-FB-UA | Oversight Board' <https://oversightboard.com/news/389395596088473-oversight-board-overturns-original-facebook-decision-case-2021-009-fb-ua/> accessed 20 September 2021

'Oversight Frameworks for Content-Sharing Platforms' (*Google*, 19 June 2019) <https://blog.google/outreach-initiatives/public-policy/oversight-frameworks-content-sharing-platforms/> accessed 5 April 2021

Pandey P, 'One Year Since the Christchurch Call to Action: A Review' (Observer Research Foundation 2020) 389

Park DH and others, 'A Literature Review and Classification of Recommender Systems Research' (2012) 39 Expert Systems with Applications 10059

'Partnering to Help Curb Spread of Online Terrorist Content' (*About Facebook*, 5 December 2016) <https://about.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/> accessed 25 March 2021

Paul C and Reininger H, 'Platforms Should Use Algorithms to Help Users Help Themselves' (20 July 2021) <https://www.rand.org/blog/2021/07/platforms-should-use-algorithms-to-help-users-help.html> accessed 27 September 2021

'Peer to Peer: Facebook Global Digital Challenge' (*EdVenture Partners*) <https://www.edventurepartners.com/peer-to-peer-facebook-global-digital-challenge> accessed 1 April 2021

'Perspective | Facebook Hides Data Showing It Harms Users. Outside Scholars Need Access.' *Washington Post* <https://www.washingtonpost.com/outlook/2021/10/05/facebook-research-data-haugen-congress-regulation/> accessed 6 October 2021

'Perspective API' <https://perspectiveapi.com/> accessed 7 September 2021

'Pittsburgh Synagogue Shooting', , *Wikipedia* (2021) <https://en.wikipedia.org/w/index.php?title=Pittsburgh_synagogue_shooting&oldid=1010099834> accessed 11 March 2021

Plantin J-C and Punathambekar A, 'Digital Media Infrastructures: Pipes, Platforms, and Politics' (2019) 41 Media, culture & society 163

'/Pol/ - On Brendon Tarrant: The Christchurch Shooter' (15 March 2019) <https://web.archive.org/web/20190315051801/https:/8ch.net/pol/res/12919462.html> accessed 11 March 2021

'Politicians, Gov't Agencies Turn to Twitter amidst #Shutdown' <https://blog.twitter.com/en_us/a/2013/politicians-govt-agencies-turn-to-twitter-amidst-shutdown.html> accessed 6 April 2021

‘Position Paper: Content Personalisation and the Online Dissemination of Terrorist and Violent Extremist Content - Tech Against Terrorism’ (17 February 2021) <https://www.techagainstterrorism.org/2021/02/17/position-paper-content-personalisation-and-the-online-dissemination-of-terrorist-and-violent-extremist-content/> accessed 27 September 2021

‘Poway Synagogue Shooting’, *Wikipedia* (2021) <https://en.wikipedia.org/w/index.php?title=Poway_synagogue_shooting&oldid=1010972146> accessed 11 March 2021

PricewaterhouseCoopers, ‘Shaping the Future of Tech Industry Regulation: Five Steps to Take Now’ (*PwC*) <https://www.pwc.com/us/en/industries/tmt/library/future-of-tech-regulation.html> accessed 7 September 2021

——, ‘The Quest for Truth: Content Moderation’ (*PwC*) <https://www.pwc.com/us/en/industries/tmt/library/content-moderation-quest-for-truth-and-trust.html> accessed 7 September 2021

‘Product Policy Forum Minutes’ (*About Facebook*, 15 November 2018) <https://about.fb.com/news/2018/11/content-standards-forum-minutes/> accessed 1 April 2021

‘Protecting Facebook Live From Abuse and Investing in Manipulated Media Research’ (*About Facebook*, 15 May 2019) <https://about.fb.com/news/2019/05/protecting-live-from-abuse/> accessed 25 March 2021

‘Protecting Users from Government-Backed Hacking and Disinformation’ (*Google*, 26 November 2019) <https://blog.google/threat-analysis-group/protecting-users-government-backed-hacking-and-disinformation/> accessed 1 April 2021

‘Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods’ (Omidyar Network and Upturn 2018) <https://omidyar.com/public-scrutiny-of-automated-decisions-early-lessons-and-emerging-methods/> accessed 6 September 2021

‘Public Statement from the Co-Chairs and European Advisory Committee of Social Science One’ <https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one> accessed 2 September 2021

‘Recommended Reading: Amazon’s Algorithms, Conspiracy Theories and Extremist Literature’ (*ISD*) <https://www.isdglobal.org/isd-publications/recommended-reading-amazons-algorithms-conspiracy-theories-and-extremist-literature/> accessed 18 September 2021

Reed A and Ingram HJ, ‘Towards a Framework for Post-Terrorist Incident Communications Strategies’ (Royal United Services Institute for Defence and Security Studies).

Reed J, ‘Soldier Kills 29 People in Thailand Shooting Rampage’ (9 February 2020) <https://www.ft.com/content/8fbe1d58-4ae7-11ea-95a0-43d18ec715f5> accessed 5 April 2021

‘Reflecting on Google’s GNI Engagement’ (*Google*, 19 December 2016) <https://blog.google/outreach-initiatives/public-policy/reflecting-googles-gni-engagement/> accessed 1 April 2021

Reliable Sources, ‘“For Teenage Girls, Is the World Better with Instagram in It or Worse?” @brianstelter Presses Facebook’s Vice President of Global Affairs Nick Clegg on Accusations That Instagram Perpetuates Body Image Issues among Its Users. Https://T.Co/TIZicv3a9w’ (*@ReliableSources*, 3 October 2021) <https://twitter.com/ReliableSources/status/1444730430461251585> accessed 4 October 2021

‘Removing Coordinated Inauthentic Behavior in UAE, Egypt and Saudi Arabia’ (*About Facebook*, 1 August 2019) <https://about.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/> accessed 25 March 2021

‘Report of the Australian Taskforce to Combat Terrorist and Extreme Violent Material Online’ (2019)

‘Report Of The Facebook Data Transparency Advisory Group’ (The Justice Collaboratory, Yale Law School 2019) <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf> accessed 8 September 2021

‘Report of the Independent International Fact-Finding Mission on Myanmar, Human Rights Council Thirty-Ninth Session 10–28 September 2018 A/HRC/39/64’ <https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf> accessed 5 October 2021

‘Research Data Agreement’ <https://socialscience.one/research-data-agreement> accessed 8 September 2021

Reuters, ‘Netizens Circumvent Moderators to Share Christchurch Shooting Video’ (*Malaysiakini*, 08:32:00+08:00) <https://www.malaysiakini.com/news/468168> accessed 18 March 2021

‘Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019’ (*Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019*, 2020) <https://christchurchattack.royalcommission.nz/> accessed 31 March 2021

‘Safety & Privacy on Twitter: A Guide for Victims of Harassment and Abuse’ <https://blog.twitter.com/en_us/a/2016/safety-privacy-on-twitter-a-guide-for-victims-of-harassment-and-abuse.html> accessed 6 April 2021

Sarah C Haan, ‘Facebook’s Alternative Facts’ (2019) 105 Virginia law review 18

Schmon C, ‘Twitter, Trump, and Tough Decisions: EU Freedom of Expression and the Digital Services Act’ (*Electronic Frontier Foundation*, 18 March 2021) <https://www.eff.org/deeplinks/2021/03/twitter-trump-and-tough-decisions-eu-freedom-expression-and-digital-services-act> accessed 8 September 2021

School SL, ‘Making Google the Censor’ (*Stanford Law School*) <https://law.stanford.edu/publications/making-google-the-censor/> accessed 6 April 2021

‘Security Council Resolution 2354 (2017) [on Implementation of the Comprehensive International Framework to Counter Terrorist Narratives]’ <http://digitallibrary.un.org/record/1298607> accessed 6 April 2021

Seetharaman GW Jeff Horwitz and Deepa, ‘Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show’ *Wall Street Journal* (14 September 2021) <https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739> accessed 15 September 2021

‘Sharing National Security Letters with the Public’ (*Google*, 13 December 2016) <https://blog.google/outreach-initiatives/public-policy/sharing-national-security-letters-public/> accessed 1 April 2021

‘She Risked Everything to Expose Facebook. Now She’s Telling Her Story.’ (*MIT Technology Review*) <https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophie-zhang-global-political-manipulation/> accessed 6 September 2021

Shead S, ‘YouTube Radicalized the Christchurch Shooter, New Zealand Report Concludes’ (*CNBC*, 8 December 2020) <https://www.cnbc.com/2020/12/08/youtube-radicalized-christchurch-shooter-new-zealand-report-finds.html> accessed 5 April 2021

‘Shedding Light on Terrorist and Extremist Content Removal’ (*RUSI*, 3 July 2019) <https://rusi.org/publication/other-publications/shedding-light-terrorist-and-extremist-content-removal> accessed 6 April 2021

Sheehy C, ‘Address Digital Harms and Rights | Global Network Initiative’ (13 October 2020) <https://globalnetworkinitiative.org/content-regulation-policy-brief/> accessed 8 September 2021

——, ‘March ’21 Submission to DSA Consultation | Global Network Initiative’ (1 April 2021) <https://globalnetworkinitiative.org/dsa-submission-mar-21/> accessed 8 September 2021

——, ‘Online Safety Bill in Australia | Global Network Initiative’ (11 May 2021) <https://globalnetworkinitiative.org/australia-online-safety-bill/> accessed 20 September 2021

Shenkman C, Thakur D and Llansó E, ‘Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis’ (Centre for Democracy and Technology 2021)

‘Shitposting, Inspirational Terrorism, and the Christchurch Mosque Massacre’ (*bellingcat*, 15 March 2019) <https://www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre/> accessed 11 March 2021

‘Significant Progress Made on Eliminating Terrorist Content Online’ (*The Beehive*) <http://www.beehive.govt.nz/release/significant-progress-made-eliminating-terrorist-content-online> accessed 5 April 2021

‘Singapore Teenager Inspired by Christchurch Massacre Arrested for Allegedly Planning Attack on Mosques, Authorities Say | The Far Right | The Guardian’ <https://www.theguardian.com/world/2021/jan/28/singapore-teenager-inspired-by-christchurch-massacre-arrested-for-allegedly-planning-attack-on-mosques-authorities-say> accessed 13 April 2021

‘Smart Regulation for Combating Illegal Content’ (*Google*, 14 February 2019) <https://blog.google/perspectives/kent-walker/principles-evolving-technology-policy-2019/smart-regulation-combating-illegal-content/> accessed 5 April 2021

‘Social Media for Social Inclusion: Tolerance and Diversity Training’ <https://blog.twitter.com/en_us/a/2015/social-media-for-social-inclusion-tolerance-and-diversity-training.html> accessed 5 April 2021

‘Some Humility About Transparency’ <http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency> accessed 12 October 2021

Staff R, ‘One Gunman, Four Locations, 29 Dead: How the Mass Shooting in Thailand Unfolded’ *Reuters* (9 February 2020) <https://www.reuters.com/article/us-thailand-shooting-timeline-idUSKBN2030FQ> accessed 5 April 2021

‘Study On The Human Rights Dimensions Of Automated Data Processing Techniques (In Particular Algorithms) And Possible Regulatory Implications’ <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5> accessed 18 September 2021

‘Supporting New Ideas in the Fight against Hate’ (*Google*, 20 September 2017) <https://blog.google/outreach-initiatives/google-org/supporting-new-ideas-fight-against-hate/> accessed 1 April 2021

‘Supporting the Vital Work of European Safety Organizations’ (*Google*, 14 May 2019) <https://blog.google/around-the-globe/google-europe/supporting-vital-work-european-safety-organizations/> accessed 1 April 2021

‘Susan Wojcicki: My Mid-Year Update to the YouTube Community’ (*blog.youtube*) <https://blog.youtube/inside-youtube/susan-wojcicki-my-mid-year-update-youtube-community/> accessed 5 April 2021

Suzor N, ‘Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms’ (2018) 4 Social media + society 205630511878781

Suzor NP, *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press 2019) <https://www.cambridge.org/core/books/lawless/8504E4EC8A74E539D701A04D3EE8D8DE> accessed 7 September 2021

——, ‘What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation’ (2019) 13 International Journal of Communication 18

‘Teaching Coding, Changing Lives: Google.Org Supports MolenGeek’ (*Google*, 5 June 2018) <https://blog.google/around-the-globe/google-europe/teaching-coding-changing-lives-googleorg-supports-molengeek/> accessed 31 March 2021

Team G, ‘Artificial Intelligence and Countering Violent Extremism: A Primer’ (*GNET*) <https://gnet-research.org/2020/09/28/artificial-intelligence-and-countering-violent-extremism-a-primer/> accessed 8 April 2021

Tech Against Terrorism, ‘The Online Regulation Series: The Handbook’ (July 2021) <https://www.techagainstterrorism.org/wp-content/uploads/2021/07/Tech-Against-Terrorism-%E2%80%93-The-Online-Regulation-Series-%E2%80%93-The-Handbook-2021.pdf> accessed 20 September 2021

‘Tech Companies Are Erasing Crucial Evidence of War Crimes’ (*Time*) <https://time.com/5798001/facebook-youtube-algorithms-extremism/> accessed 9 April 2021

‘Text-Driven Normativity’ (*COHUBICOL publications*, 30 July 2021) <https://publications.cohubicol.com/working-papers/text-driven-normativity/> accessed 27 September 2021

‘Thai Commandos Kill Rogue Soldier Who Shot Dead 29 People’ <https://www.aljazeera.com/news/2020/2/9/thai-commandos-kill-rogue-soldier-who-shot-dead-29-people> accessed 5 April 2021

‘Thailand Shooting: Soldier Who Killed 26 in Korat Shot Dead’ *BBC News* (9 February 2020) <https://www.bbc.com/news/world-asia-51431690> accessed 5 April 2021

‘The Christchurch Attack Changed How Counter-Terrorism Thinks about Online Propaganda, Hashing and the Role of AI’ (*Faculty*) <https://faculty.ai/blog/the-christchurch-attack-changed-how-counter-terrorism-thinks-about-online-propaganda-hashing-and-the-role-of-ai/> accessed 5 April 2021

‘The Christchurch Attacks: Livestream Terror in the Viral Video Age’ (*Combating Terrorism Center at West Point*, 18 July 2019) <https://ctc.usma.edu/christchurch-attacks-livestream-terror-viral-video-age/> accessed 14 April 2021

‘The Digital Services Act and Freedom of Expression: Triumph or Failure? - Blog - Maastricht University’ <https://www.maastrichtuniversity.nl/blog/2021/03/digital-services-act-and-freedom-expression-triumph-or-failure> accessed 8 September 2021

‘The Essential Tech Worker’ (*POLITICO*, 19 October 2020) <https://www.politico.eu/special-report/the-essential-tech-worker-coronavirus-public-health-pandemic/> accessed 7 September 2021

‘The EU Code of Conduct on Countering Illegal Hate Speech Online’ (*European Commission - European Commission*) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 1 April 2021

‘The Facebook Files - Yale Law School’ <https://law.yale.edu/isp/events/facebook-files> accessed 8 October 2021

‘The Facebook Files, Part 1: The Whitelist - The Journal. - WSJ Podcasts’ (*WSJ*) <https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-1-the-whitelist/72a1e8f5-a187-4a91-bedb-b0b0d39f5cce> accessed 29 September 2021

‘The Facebook Files, Part 2: “We Make Body Image Issues Worse” - The Journal. - WSJ Podcasts’ (*WSJ*) <https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-2-we-make-body-image-issues-worse/c2c4d7ba-f261-4343-8d18-d4de177cf973> accessed 29 September 2021

'The Facebook Files, Part 3: "This Shouldn't Happen on Facebook" - The Journal. - WSJ Podcasts' (*WSJ*) <https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-3-this-shouldnt-happen-on-facebook/0ec75bcc-5290-4ca5-8b7c-84bdce7eb11f> accessed 29 September 2021

'The Facebook Files, Part 4: The Outrage Algorithm - The Journal. - WSJ Podcasts' (*WSJ*) <https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f> accessed 29 September 2021

'The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why.' (*MIT Technology Review*) <https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/> accessed 6 October 2021

'The Four Rs of Responsibility, Part 1: Removing Harmful Content' (*blog.youtube*) <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/> accessed 1 April 2021

'The German Synagogue Shooter's Twitch Video Didn't Go Viral. Here's Why.' <https://www.vice.com/en/article/zmjgzw/the-german-synagogue-shooters-twitch-video-didnt-go-viral-heres-why> accessed 6 April 2021

'The Great Replacement : Featured Classification Decisions : OFLC' <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/the-great-replacement/> accessed 18 March 2021

'The Hate-Filled Website 8chan Was Taken Offline After the El Paso Shooting. By Monday Morning It Was Back.' <https://www.vice.com/en/article/mbmqpp/the-hate-filled-website-8chan-was-taken-offline-after-the-el-paso-shooting-by-monday-morning-it-was-back> accessed 6 April 2021

'The Lawfare Podcast: Ben Smith on Gatekeepers in the Internet Age' (*Lawfare*, 11 February 2021) <https://www.lawfareblog.com/lawfare-podcast-ben-smith-gatekeepers-internet-age> accessed 5 April 2021

'The Lawfare Podcast: Canada Takes on the Proud Boys' (*Lawfare*, 12 February 2021) <https://www.lawfareblog.com/lawfare-podcast-canada-takes-proud-boys> accessed 5 April 2021

'The Lawfare Podcast: Content Moderation and the First Amendment for Dummies' (*Lawfare*, 11 March 2021) <https://www.lawfareblog.com/lawfare-podcast-content-moderation-and-first-amendment-dummies> accessed 5 April 2021

'The Lawfare Podcast: Facebook Shuts Down Research On Itself' (*Lawfare*, 19 August 2021) <https://www.lawfareblog.com/lawfare-podcast-facebook-shuts-down-research-itself> accessed 2 September 2021

'The Lawfare Podcast: Jacob Schulz on Seditious Conspiracy' (*Lawfare*, 24 March 2021) <https://www.lawfareblog.com/lawfare-podcast-jacob-schulz-seditious-conspiracy> accessed 5 April 2021

'The Lawfare Podcast: Russia Cracks Down on Social Media' (*Lawfare*, 7 October 2021) <https://www.lawfareblog.com/lawfare-podcast-russia-cracks-down-social-media> accessed 10 October 2021

'The Lawfare Podcast: Tech CEOs Head to the Hill, Again' (*Lawfare*, 1 April 2021) <https://www.lawfareblog.com/lawfare-podcast-tech-ceos-head-hill-again> accessed 5 April 2021

'The Lawfare Podcast: Trust, Software and Hardware' (*Lawfare*, 22 February 2021) <https://www.lawfareblog.com/lawfare-podcast-trust-software-and-hardware> accessed 5 April 2021

'The Lawfare Podcast: YouTube, We Have a Problem' (*Lawfare*, 25 March 2021) <https://www.lawfareblog.com/lawfare-podcast-youtube-we-have-problem> accessed 5 April 2021

‘The New Tool Helping Asian Newsrooms Detect Fake Images’ (*Google*, 25 February 2020) <https://blog.google/around-the-globe/google-asia/new-tool-helping-asian-newsrooms-detect-fake-images/> accessed 1 April 2021

‘The Redirect Method’ <http://redirectmethod.org> accessed 1 April 2021

Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019 <https://christchurchattack.royalcommission.nz/the-report/> accessed 31 March 2021

‘The Root Causes of Violent Extremism, 04 January 2016’ <https://ec.europa.eu/home-affairs/orphan-pages/page/root-causes-violent-extremism-04-january-2016_sl> accessed 27 September 2021

‘The Secret Language of Fans’ <https://blog.twitter.com/en_us/topics/insights/2019/the-secret-language-of-fans.html> accessed 6 April 2021

‘The Twitter Rules: Safety, Privacy, Authenticity, and More’ <https://help.twitter.com/en/rules-and-policies/twitter-rules> accessed 6 September 2021

‘They Are Us’ <http://shorthand.radionz.co.nz/together-alone/index.html> accessed 11 March 2021

‘This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook’ (*BuzzFeed News*) <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> accessed 24 March 2021

‘Thread by @RebekahKTromble on Thread Reader App’ <https://threadreaderapp.com/thread/1444704706597642250.html> accessed 4 October 2021

‘Three Lessons in Content Moderation from New Zealand and Other High-Profile Tragedies - Center for Democracy and Technology Three Lessons in Content Moderation from New Zealand and Other High-Profile Tragedies - Center for Democracy and Technology’ <https://perma.cc/436Z-Z6J7> accessed 6 April 2021

‘Tips for Engaging Live: How Automakers Used Periscope at #NYIAS’ <https://blog.twitter.com/en_us/a/2016/tips-for-engaging-live-how-automakers-used-periscope-at-nyias.html> accessed 5 April 2021

‘To Stop Terror Content Online, Tech Companies Need to Work Together’ (*Google*, 20 December 2018) <https://blog.google/outreach-initiatives/public-policy/stop-terror-content-online-tech-companies-need-work-together/> accessed 5 April 2021

‘Toomas Hendrik Ilves: Is Social Media Good or Bad For Democracy?’ (*About Facebook*, 25 January 2018) <https://about.fb.com/news/2018/01/ilves-democracy/> accessed 24 March 2021

Tuesday and others, ‘Shared Crisis Response Protocol | Scoop News’ <https://www.scoop.co.nz/stories/PA1912/S00014/shared-crisis-response-protocol.htm> accessed 8 March 2021

Twitter, ‘Filter Realtime Tweets: Overview (Streaming API)’ <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview> accessed 8 September 2021

——, ‘Digital Services Act: Defending the Open Internet’ <https://blog.twitter.com/en_us/topics/company/2021/the-digital-services-act--defending-the-digital-single-market> accessed 8 September 2021

‘Twitter and PBS NewsHour Partner to Live Stream Coverage of Inauguration’ <https://blog.twitter.com/en_us/topics/events/2017/twitter-and-pbs-newshour-partner-to-live-stream-coverage-of-inauguration-day-2017.html> accessed 5 April 2021

'Twitter Recently Joined Young People in Vienna to Talk about Alternative Narratives & Changing Attitudes' <https://blog.twitter.com/en_us/a/2016/twitter-recently-joined-young-people-in-vienna-to-talk-about-alternative-narratives-changing.html> accessed 6 April 2021

'Twitter Supports Radicalisation Awareness Network Campaign Encouraging Europeans to #ExitHate' <https://blog.twitter.com/en_us/a/2016/twitter-supports-radicalisation-awareness-network-campaign-encouraging-europeans-to-exithate.html> accessed 6 April 2021

'Twitter's Decentralized Future' (*TechCrunch*) <https://social.techcrunch.com/2021/01/15/twitters-vision-of-decentralization-could-also-be-the-far-rights-internet-endgame/> accessed 11 March 2021

UN High Commissioner for Human Rights, Michelle Bachelet, 'Letter from the UN High Commissioner for Human Rights to the President of the European Commission Regarding the Digital Services Act Proposal' (7 September 2020) <https://europe.ohchr.org/EN/Stories/Documents/2020%2009%2007%20Letter%20HC%20to%20EC%20President.pdf> accessed 8 September 2021

'#UNGA: Twitter and the Global Political Conversation' <https://blog.twitter.com/en_us/a/2013/unga-twitter-and-the-global-political-conversation.html> accessed 5 April 2021

'Update on New Zealand' (*About Facebook*, 19 March 2019) <https://about.fb.com/news/2019/03/update-on-new-zealand/> accessed 25 March 2021

'Update on Our Advertising Transparency and Authenticity Efforts' (*About Facebook*, 27 October 2017) <https://about.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/> accessed 25 March 2021

'Update on the Global Internet Forum to Counter Terrorism' (*Google*, 4 December 2017) <https://blog.google/around-the-globe/google-europe/update-global-internet-forum-counter-terrorism/> accessed 5 April 2021

'Update on User Safety Features' <https://blog.twitter.com/en_us/a/2015/update-on-user-safety-features.html> accessed 6 April 2021

'Using Data to Change the Conversation about Race in America' (*Google*, 13 June 2017) <https://blog.google/outreach-initiatives/google-org/using-data-change-conversation-about-race-america/> accessed 5 April 2021

Victoria Jordan, Kristin Thue, and Jacopo Bellasio, 'Malign Use of Algorithmic Amplication of Terrorist and Violent Extremist Content: Risks and Countermeasures in Place' (European Commission 2021)

Vincent J, 'Archive.Org Hit with Hundreds of False Terrorist Content Notices from EU' (*The Verge*, 11 April 2019) <https://www.theverge.com/2019/4/11/18305968/eu-internet-terrorist-content-takedown-mistakes-internet-archive-org> accessed 8 September 2021

——, 'Facebook Is Now Using AI to Sort Content for Quicker Moderation' (*The Verge*, 13 November 2020) <https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation> accessed 13 April 2021

'Violent White Supremacy' (*Jigsaw*) <https://jigsaw.google.com/the-current/white-supremacy/> accessed 1 April 2021

Techdirt 'Podcast: Regulating Amplification Is A Lot Harder Than You Think [Ep.287]' (*Techdirt.*) <https://www.techdirt.com/articles/20210623/12462247046/techdirt-podcast-episode-287-regulating-amplification-is-lot-harder-than-you-think.shtml> accessed 27 September 2021

Wegener F, 'How the Far-Right Uses Memes in Online Warfare' (*GNET*) <https://gnet-research.org/2020/05/21/how-the-far-right-uses-memes-in-online-warfare/> accessed 13 April 2021

——, 'The Globalisation of Right-Wing Copycat Attacks' (*GNET*) <https://gnet-research.org/2020/03/16/the-globalisation-of-right-wing-copycat-attacks/> accessed 13 April 2021

Weltorganisation für Geistiges Eigentum (ed), *Intellectual Property and Human Rights: A Panel Discussion to Commemorate the 50th Anniversary of the Universal Declaration of Human Rights, Geneva, November 9, 1998* (Reprint, World Intellectual Property Organization 2000)

'What Are My Ad Topic Preferences and How Can I Adjust Them on Instagram? | Instagram Help Centre' <https://www.facebook.com/help/instagram/245100253430454> accessed 6 September 2021

'What Happened When Humans Stopped Managing Social Media Content' (*POLITICO*, 21 October 2020) <https://www.politico.eu/article/facebook-content-moderation-automation/> accessed 7 September 2021

'What Is the Content Incident Protocol?' (*GIFCT*) <https://gifct.org/?faqs=what-is-the-content-incident-protocol> accessed 18 March 2021

'What Our Research Really Says About Teen Well-Being and Instagram' (*About Facebook*, 26 September 2021) <https://about.fb.com/news/2021/09/research-teen-well-being-and-instagram/> accessed 27 September 2021

'What to Expect on Twitter on US Inauguration Day 2021' <https://blog.twitter.com/en_us/topics/company/2021/inauguration-2021.html> accessed 5 April 2021

Whittaker J and others, 'Recommender Systems and the Amplification of Extremist Content' (2021) 10 Internet Policy Review <https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content> accessed 8 September 2021

'Widows of Shuhada' (*RNZ*) <https://www.rnz.co.nz/programmes/widows-of-shuhada> accessed 11 March 2021

Won YB, 'Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts' (*GNET*) <https://gnet-research.org/2021/04/06/male-supremacism-borderline-content-and-gaps-in-existing-moderation-efforts/> accessed 8 April 2021

'Working Together to Combat Terrorists Online' (*Google*, 20 September 2017) <https://blog.google/outreach-initiatives/public-policy/working-together-combat-terrorists-online/> accessed 1 April 2021

'World Leaders on Twitter: Principles & Approach' <https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html> accessed 5 April 2021

'Writing Facebook's Rulebook' (*About Facebook*, 10 April 2019) <https://about.fb.com/news/2019/04/insidefeed-community-standards-development-process/> accessed 1 April 2021

York JC, 'UN Report Sets Forth Strong Recommendations for Companies to Protect Free Expression' (*Electronic Frontier Foundation*, 27 June 2018) <https://www.eff.org/deeplinks/2018/06/un-report-sets-forth-strong-recommendations-companies-protect-free-expression> accessed 7 September 2021

'YouTube's Recommender AI Still a Horror Show, Finds Major Crowdsourced Study' (*TechCrunch*) <https://social.techcrunch.com/2021/07/07/youtubes-recommender-ai-still-a-horrorshow-finds-major-crowdsourced-study/> accessed 8 September 2021

Zannettou S and others, 'On the Origins of Memes by Means of Fringe Web Communities' [2018] arXiv:1805.12512 [cs] <http://arxiv.org/abs/1805.12512> accessed 6 April 2021

Zerilli J and others, 'Algorithmic Decision-Making and the Control Problem' <https://www.repository.cam.ac.uk/handle/1810/314956> accessed 29 September 2021

——, 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?' <https://www.repository.cam.ac.uk/handle/1810/299973> accessed 29 September 2021