ISD | Powering solutions to extremism and polarisation

Sponsored by:
Federal Foreign Office

Digital Policy Lab '20 Companion Papers

- Transparency, Data Access and Online Harms
- National & International Models for Online Regulation
- Policy Summary: EU Digital Services Act & UK Online Safety Bill
- The Liberal Democratic Internet – Five Models for a Digital Future
- Future Considerations for Online Regulation

Chloe Colliver, Milo Comerford, Jennie King, Alex Krasodomski-Jones, Christian Schwieter & Henry Tuck

About the Digital Policy Lab

The Digital Policy Lab (DPL) is a new inter-governmental working group focused on charting the regulatory and policy path forward to prevent and counter disinformation, hate speech, extremism and terrorism online.

It is comprised of a core group of senior representatives of relevant ministries and regulators from key liberal democratic countries. The DPL is intended to foster inter-governmental exchange, provide policy makers with access to sector-leading expertise and research, and build an international community of policy practice around key regulatory challenges in the digital policy space.

The project is funded by the German Federal Foreign Office.

The views expressed in these papers are those of the authors and do not necessarily reflect the views of Digital Policy Lab participants or funders.

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

ISD | Powering solutions to extremism and polarisation

Digital Policy Lab
Discussion Paper

Transparency, Data Access and Online Harms

About This Paper

This discussion paper provides an overview of the evolution of the international policy debate around data access and transparency to counter disinformation, hate speech, extremism and terrorism online. It is intended as an overview of the key issues covered by the first Digital Policy Lab event on 12–13th November 2020, and incorporates discussions from the event.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of Digital Policy Lab participants or governments.

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Why does transparency matter?

The migration of our lives onto the internet has created powerful new forces that shape citizens' lives. Online platforms have radically altered the ways in which we communicate, make decisions and come to hold views. At the same time, governments across the globe have lamented the emergence of a diverse set of online harms. The explosion of the Covid-19 'infodemic' throughout 2020 has thrown into sharp relief the range of online harms facing liberal democratic societies, from disinformation to hate speech, extremism, and terrorism, and demonstrates the urgency of international coordination to mitigate the threat.

However, unless there is an informed understanding of the scale and nature of these challenges, it is incredibly difficult to understand which potential solutions would be both effective and proportional. It is of central importance that governments, regulators, civil society and the public at large are able to better understand the ways in which the internet is impacting society and democracy in order to enhance its positive effects, and to limit negative externalities.

Governments and elected officials require better information to fulfil their obligations to citizens, including upholding and enforcing existing laws, providing democratic oversight, protecting national security, representing the perspectives of constituents that have been the victims of online harms, and advocating for change on their behalf. Similarly, for regulators to be effective, they need a more complete understanding of company policies, procedures and decisions, as well the underlying technology, its outputs and potential biases. Civil society, academia and the media would benefit from greater access to data to fulfil their public interest mandates and illuminate the often opaque online world. This would build the evidence-base on the perpetrators, causes and impacts of online harms, and provide independent scrutiny, advice, and support to vulnerable or minority groups. Finally, transparency is vital for the public to understand their rights and responsibilities online, the relationships they enter into with online platforms, and the environment in which they spend increasing amounts of time, receive their information, and participate in society and the economy.

In order to effectively provide oversight and develop sustainable policies, legislation and regulation for the online world, we will require a more substantive evidence base. Transparency is the lever through which this evidence can be gathered.

Transparency in its broadest sense provides a mechanism for improving visibility, understanding and accountability for public policy issues. By increasing transparency of online spaces and platforms, the argument goes, we stand a better chance of detecting, mitigating and responding to this broad spectrum of both illegal and legal online harms. Transparency has been widely cited and accepted as a key principle for 'good governance' of public administration, including the Council of Europe, the OSCE and the European Commission. In short, it is assumed that, in order for governance to be fair and efficient, independent oversight (either regulatory or non-regulatory) and avenues for public scrutiny are necessary. In a democratic system, transparent processes need to be in place that ensure public actors can be held accountable.

ISD's decade-long experience working with the private sector, policy makers and civil society across a range of online challenges, from terrorism to hate speech and disinformation, shows that any effective regulatory or non-regulatory approaches to tackling this spectrum of online harms should be rooted in transparency. Transparency is not an end in itself, but a prerequisite to establish public trust, accountability, oversight, and a healthy working relationship between tech companies, government, and the public. However, the requirements and expectations associated with transparency are often poorly articulated, lack specificity, or vary across online platforms and offline jurisdictions. Promising frameworks and models for transparency exist, both in the digital context and other related or comparable areas, and should be used as best practice when considering the wider online ecosystem and the future of expectations for transparency online.

Balancing transparency with safety and data protection concerns

Efforts to address harmful activity online have been described as an arms race between malign actors and those seeking to counter them. This context requires an effective balance, as transparency of enforcement approaches might be a disadvantage in the longer term. However, despite this difficult and evolving information environment, the highest level of transparency is nonetheless vital for building trust, ensuring rigour, and fighting abuse. Concerns about malign actors understanding the precise methods used for detection and moderation are valid, but should not necessarily preclude the development of accountability structures to address concrete vulnerabilities. Furthermore, it is important to note that transparency must complement rights to data privacy, not erode them. A good model for transparency will protect individuals' data privacy while enabling a macro understanding of the nature and scale of technology platforms' processes and any potential infringement of rights that stems from the use of the platform.

Challenges posed by a lack of transparency

Limitations of current public research on online harms

From a civil society perspective, ISD's own research has shown the limitations of transparency when understanding harmful behaviour online. A recent study reviewing the action taken by Facebook against 'coordinated inauthentic behaviour' – defined as organised covert, deceptive and deliberately misleading activity on the platform – revealed that even public transparency reports from tech companies tell only a partial story. Crucially, in their current format they do not enable a deep or holistic understanding of key issues, either in terms of scale, nature or impact. What information there is available shows the significant scale of deceptive activity targeting electorates around the globe on Facebook, from nation states, public relations companies and ideologically motivated hate groups, among others. What it does not and cannot tell us is the true scale of this kind of activity, including that which is not detected or reported by the company. Independent researchers, including from ISD, continue to identify examples of large-scale 'coordinated inauthentic behaviour' on the platform, despite having minimal access to Facebook data. In the lead-up to the European Parliamentary elections for instance, ISD and partner organisations identified nineteen such coordinated inauthentic networks on Facebook and Twitter through research focused just on six EU member states. The evidence suggests that the examples provided by Facebook over the past two years only represent a fraction of the true scale of such activity on the platform. This has only been further exacerbated in the wake of the Covid-19 'infodemic'. An investigation by ISD and the BBC found a set of websites spreading disinformation around COVID-19 had received over 80 million engagements on Facebook during the health crisis, six times the combined engagement for the US Centers for Disease Control and Prevention (CDC) and World Health Organisation (WHO).

In a similar vein, recent ISD research on the scale of online abuse targeting a variety of politicians, found that women and candidates from an ethnic minority background tended to be disproportionately abused online. This research demonstrated the need for social media platforms to provide greater transparency about their content moderation policies, processes and enforcement outcomes relating to harassment and abuse. This included the type of content that fallswithin and beyond the remit of their relevant policies, the resources allocated to content moderation, the linguistic and cultural contextual expertise of those teams, and the appeals and redress processes in place for wrongful removals.

Recently, increased scrutiny has focused on rapidly growing youth-focused social media platforms such as TikTok, where a number of potential harms have been flagged, including public health disinformation during the Covid-19 pandemic. While the platform has pledged to hold itself accountable to its community through transparency, by sharing information about content removal, including hate speech and misinformation, the lack of an official API (Application Programme Interface), opaque search functions and features like the mobile-first design of the platform make it extremely hard for researchers to automate data collection or view trends at scale.

While there have been private sector attempts to provide greater access to academic researchers, there are undoubtedly challenges in developing safe and transparent processes for anonymised data-sharing with third parties at scale. Initiatives such as Social Science One, designed to provide vetted academics with access to large anonymised datasets to study disinformation on Facebook, are a positive step forward, but still relatively narrow in scope and yet to deliver major results. This only reinforces the fact that limited co-operation between companies and independent researchers is hampering progress in this area.

Revelations from tech companies revealing the lack of transparency

Alongside insights gleaned from company transparency reports, and the variable levels of access for researchers provided by online platforms, information extracted from companies via democratic oversight (or leaked to the media) demonstrates the extent to which we cannot fully understand the online environments in which online harms thrive, and the decisions of private companies that shape them. An indication of the scale of this challenge is provided by insights from tech company insiders, as well as revelations about the internal information available to platforms which

are not shared with governments, civil society or the public through more limited approaches to transparency reporting.

For example, in September 2020, a recently fired Facebook employee, Sophie Zhang, wrote a memo criticising Facebook for failing to respond effectively to global inauthentic and coordinated political activity on the platform. She raised concerns that researchers and policymakers have highlighted for some time, namely that Facebook enabled political operatives all over the world to conduct deceptive activity targeting elections at enormous scale, with a very low bar for entry. Ironically, the ability of researchers and civil society to expose such activities was limited after Facebook significantly reduced third-party API access over recent years, and in particular after the Cambridge Analytica scandal. While this was intended to prevent abuse by commercial or political actors, it has also stymied efforts to research online harms. For example, the company has recently sought the removal of a New York University tool designed to increase the transparency of political advertising targeting. In contrast, documents previously obtained by the UK House of Commons illustrated how data access for some select private companies was expanded as part of special 'whitelisting' agreements.

There is also limited transparency of the analytical capabilities and data available internally to platforms, beyond the simple metrics around content takedown presented publicly in transparency reporting. For example, longstanding concerns around recommendation algorithms have proven difficult to substantiate by academics or civil society researchers, who have only access to a small subset of data. However, a recently leaked internal Facebook report presented to executives in 2018 found that the company was well aware that its product, specifically its recommendation engine, stoked divisiveness and polarization. Indeed as early as 2016 an internal report found that 64% of people who joined an extremist group on Facebook only did so because the platform's algorithm recommended it to them, according to the Wall Street Journal.

Other studies have leveled similar criticism at YouTube's recommendation algorithm and its role in facilitating political polarisation and radicalisation. However, a lack of access to data for independent researchers has also made it near impossible to verify YouTube's claims that it has reduced recommendations of "borderline content and harmful misinformation" by 50%. The Mozilla Foundation has laid out a number of proposals for garnering meaningful data for researchers, including richer impression and engagement metrics, access to historical video archives, and simulation tools which allow for better understanding of recommendation algorithm pathways.

Acknowledging this potential harm of recommendation systems, Facebook quietly suspended political group recommendations ahead of the 2020 US election, when it was also revealed that the company had established a metric for monitoring 'violence and incitement trends'. This tool, which assesses the potential for danger based on hashtags and search terms – and which purported to have found a 45% increase in violence and incitement over the election period – demonstrates the considerable potential for increased real time data which is potentially available to private companies, but not accessible to governments, regulators or researchers. It remains unclear how such tools relate to or are used to inform or trigger policy and enforcement responses by the company, including whether a certain level of 'violence and incitement' is deemed acceptable.

Key areas requiring further transparency

In a 2019 paper, ISD laid out four key areas of technology companies' and platforms' functions and services that require improved transparency to better address online harms and uphold rights:

Content & Moderation

Platforms that have become public spaces must make that space as intelligible as possible. As web platforms and their users play an increasing role in shaping our culture, informing our political decision-making, and driving societal change, the activities taking place in these spaces should be observable. Transparency here calls for both researchers and users to be provided access to public content in a systematic way, as well as clear information about how platforms moderate this content.

The former can be achieved through an easily-navigable API, giving the public a means to query live and historical content beyond the constraints of the default 'news feed'. Beyond the official API offered by Twitter and other services such as Facebook-owned CrowdTangle, some researchers have developed their own independent monitoring and search capabilities to improve the visibility of platform activity.

Beyond API access, recent regulatory initiatives in Germany, France and Australia that require platforms to take quicker action on illegal or hateful content have emphasised the need for greater transparency when it comes to platform moderation activities. Many of these initiatives require social media companies to provide regular transparency reports documenting complaints received and decisions taken against hate speech or coordinated inauthentic behaviour (CIB) on their platforms. In addition, many constituencies require social media companies to publicise information or content-blocking requests by law enforcement.

These types of transparency reports can be an important tool to support researchers, civil society, regulators and policymakers in their work. For example, a recent ISD investigation has used the publicly available archive of Facebook's transparency reports to both understand how the company identifies and combats inauthentic activity on its platforms, and to unveil the significant revenue Facebook was able to generate from these accounts in the form of ad sales. Additionally, they can provide clarity on the evolving terms of services and their enforcement on behalf of platforms.

Complaints & Redress

A significant gap exists in the public's understanding of platforms' ability to moderate and respond to abuses of their platforms. Visibility of complaints made to platforms is essential to accountability, to support the victims of online harms, to raise awareness of challenges facing users online, and to provide evidence in redress. As described above, regular transparency reports, sometimes legally mandated, have sought to fill this gap in public understanding.

However, transparency reports have often failed to provide meaningful insight into the moderation processes of private companies, therefore limiting the ability of users to appeal and challenge decisions. In fact, the first fine levied under the German NetzDG law targeted Facebook for providing incomplete data in its first 2018 transparency report. In response to mounting pressure, Facebook announced the creation of an 'independent oversight board'. After its 40 board members were introduced in May 2020, the board began reviewing cases in late 2020. The goal is to allow users to appeal content decisions made by Facebook by escalating them to the board members, whose decisions are binding and will shape Facebook's moderation policies going forward.

A similar oversight role is currently played by the German Association for Voluntary Self- Regulation of Digital Media Service Providers (FSM e. V.). Mandated as the 'regulated self-regulator' under the NetzDG law, social media companies can decide to delegate difficult content decisions to expert review panels convened by the FSM. The FSM also has procedures in place to review user appeals against content moderation decisions on behalf of the social media company. Since April 2020, ten such user complaints have been received, of which six were deemed to be legal content and hence had to be reinstated by the social media companies.

Advertising

Advertising - particularly targeted political advertising - is one of the core products offered by many online platforms, allowing advertisers and campaigners to deliver content directly to a chosen audience. It is in the public interest for internet users to understand how and why they are being targeted online, and for regulators to be able to understand and respond to malpractice.

Many constituencies around the world have determined that transparency requirements for political advertising ought to go beyond those expected of unpaid or organic public content and communications. Regulatory initiatives, such as those in France, Ireland, Australia, Canada, the US, the UK and the EU, have proposed the expansion of existing authorisation requirements for offline political advertising to the online realm. This not only includes requirements for clear labelling on paid-for content noting the address of who authorised the ad, but also providing users with information about why they were targeted.

To meet these demands, Facebook and Twitter have introduced a public archive of ads that can be explored and queried by anyone. However, the lack of details provided and unreliability of the service during key election phases have shown the importance of civil society-led initiatives such as the UK-based Who Targets Me or the aforementioned NYU Ad Observatory to improve online ad transparency. Many of these initiatives pre-date the official platform ad archives and use browser plug-ins to crowdsource information about where and when ads pop up on users' news feeds.

Platform Architecture & Design

There remains significant concern that platform architectures contribute to negative societal outcomes. These range from evidence about racial and gender bias in search engine results to worries that platforms' user interface and design incentivises the spread of divisive or misleading content. Central to these concerns is that the algorithms dictating a users' experience and user journey have led to unintended consequences, and have been challenging to scrutinise or evaluate for those unfamiliar with the internal operations of social media companies. For example, recommendation systems have been criticised for driving users to consume ever more extreme content, potentially facilitating political radicalisation. This is particularly notable for YouTube, whose representatives have claimed that 70% of the time spent on YouTube is driven by recommended videos, rather than organic user searches.

While there are some attempts by social media companies to provide transparency on recommendation algorithms, these are often incomprehensible to the average user. For researchers, recommendation systems make it difficult to gain an objective understanding of the reach of content: as the system adapts to each user, each individual news feed, and hence exposure to content, is unique. The curation decisions and user signals that determine what content is presented remain opaque to outsiders. Facebook for example recently refuted claims that conservative commentators have a wider reach on its platform than traditional media outlets. Journalists had used Facebook's publicly available data to aggregate reactions and comments per day and week to show how, on an average day, the most popular content is dominated by conservative pundits. In a blog post, Facebook stated that engagement figures and reach are two different metrics - however, the latter metric is not available to independent researchers or the wider public.

Similar issues have been raised by the French Conseil supérieur de l'audiovisuel (CSA), where a lack of transparency by platforms regarding their curation algorithms has limited the ability of the regulator to perform proper oversight. In terms of further statutory regulation, a recent report by the Ada Lovelace Institute has published detailed guidance on how different types of algorithmic audits and impact assessment can be leveraged to improve platform transparency and ultimately company accountability in this area.

Approaches to Achieving Greater Transparency

Additional transparency is required in terms of the policies and processes that apply in each of these four areas, and the outcomes and impacts that they produce. Companies' policies, defined through their Terms of Service, determine the rules that are in place across their platforms. It is therefore vital to understand the evidence base used to construct these rules, and who was responsible for such decisions. Companies' processes then determine how these policies are enacted and enforced in practice, including the human and automated systems that make decisions on a daily basis. Again, transparency is vital to better understand how these processes have been designed and by whom, and what safeguards are in place to ensure consistency and prevent biases from occurring. Finally, further transparency will help determine whether decision-making frameworks are being driven by efforts to improve user safety and experience, or the commercial and business incentives of the company itself. To address these needs, a variety of models have been developed in response to the emergence of online harms associated with social media platforms and digital services more broadly. These models differ in terms of their intended scope, with some focused on specific aspects of platforms' functions and services, and/or the policies, processes and outcomes in these areas, while others attempt to adopt a more holistic approach that encompasses several or all of these different areas and approaches. Below is a brief overview illustrative of these different but overlapping models.

Procedural accountability

The 'French framework to make social media platforms more accountable', published in May 2019, emphasises the need for close cooperation between private companies, an independent regulator and the government. The point of departure here is that recent years have seen a deterioration of trust between social media companies, regulators and users. At the heart of this is an asymmetry of information that can only be improved through an ongoing open dialogue between companies, regulators and civil society. Transparency, then, is a necessary first step for the whole of society to be able to address the issue of online harms.

In contrast to ex-post regulation via notice-and-takedown regimes, the French framework calls for improved procedural accountability, that is "imposing strong obligations on the transparency of key systems unobservable from the outside, i.e. the moderation system (and procedures for developing and updating the terms of use that underlies it), as well as the use of algorithms for targeting and personalising the content presented." Borrowing from the financial audit model of the banking industry, the framework proposes an independent regulator who would primarily ensure 'systemic' companies are implementing preventive and corrective measures. The regulator would not seek to regulate specific cases where harm may materialise or has materialised, but instead impose a 'duty of care' for social media companies that are deemed 'systemic' actors.

Transparency by default

The UK Information Commissioner's Office (ICO) code of practice for age-appropriate design adopts a different approach to transparency and accountability, focusing on the 'user-facing' side alongside company-internal processes. At the heart of the code is the assumption that transparency is primarily about communicating with users in a 'clear, open and honest' manner about 'what they can expect when they access online services'. Aside from clear descriptors about how personal data may be used, the code also calls for transparent and easily-understandable terms of service and community standards. In addition, 'nudging' is explicitly mentioned as a tool that can be used proactively to encourage users to be more privacy-conscious, for example by opting out of recommendation systems.

The lack of user-friendly design is also criticised in the recent independent review of the German NetzDG law and has been a recurring point raised by Jugendschutz.net - both call for more transparent and easily accessible reporting procedures on behalf of social media companies. This includes an easily visible reporting button to help users, including children and adolescents, flag inappropriate content they may come across on platforms. The core rationale of this 'safety-by-design' approach is that online harms do not only materialise through content, but online interaction between users mediated by social media companies.

Both the ICO's code and the work of Jugenschutz.net primarily seek to protect minors from online harms, levying user experience research to promote privacy-conscious and safe online behaviour - this could equally be adopted as a model to hold digital service providers accountable in other settings, including social media use by adults.

A human right's framework for digital policymaking

Among others, David Kaye, the former United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, has advocated for a rights-based approach to online regulation and particularly content moderation. His central claim is that human rights standards as set out in the Universal Declaration of Human Rights should guide content moderation norms. This does not mean that any form of expression should be allowed on social media - rather, company's terms of services and government regulation should clearly articulate when and why restricting the right to freedom of expression is necessary and proportionate.

The attempt to regulate only where it is necessary lies at the heart of the risk-based approach foreseen by the UK Online Harms White Paper, or the focus on systemic actors by the French framework. However, Kaye and the Forum for Democracy & Information go further by demanding that, in their decisions, companies or regulators should refer to international human rights jurisprudence. This necessitates 'rulemaking transparency' as well as 'decisional transparency' by clearly laying out the decision-making process behind platform action. This transparency can then provide the basis for company and government accountability as the public can scrutinise, and appeal, decisions made.

Based on this approach, the recent report by the Forum for Information & Democracy has further proposed that the only ex-ante requirement of platforms in terms of transparency should be a so-called human rights impact assessment of their services and any proposed changes to their services, including content moderation practices.

Annex: Transparency Framework

The following framework for understanding where additional transparency may be required was presented by ISD during the DPL event in November 2020. It combines the various categories outlined in this briefing, and provides selected examples of current questions relating to online harms, existing initiatives (both public and private sector-led), as well as areas where additional transparency may be required.

Examples of Online Harms	Policies	Processes	Outcomes
Disinformation Study 1 Study 2 Study 3 Study 4	What constitutes illicit coordination? How do companies differentiate between fake and coordinated accounts?	How comprehensive is the linguistic and cultural scope of these investigations and teams? Are Iran and Russia the most cited sources because that is where companies are looking?	What is the reach of coordinated networks into real user conversations on social media? How many users engaged with content disseminated by inauthentic accounts?
Conspiracy theories & extremist recruitment Study 1 Study 2 Study 3 (German) NBC report	Why are some extremist groups allowed to remain active online and others are removed? Who makes those decisions?	How do recommendation algorithms promote or demote extremist groups? How did the platform's internal targets to grow 'group' membership proactively impact how many users joined QAnon groups?	What is the number of members of public and private extremist groups? What are the profits generated by advertising purchased by extremist groups?
Harassment and abuse Study 1	What type of content (incl. text, image, video) falls inside and outside of relevant abuse policies? What are company policies on retaining evidence of harassment and threats to life for legal recourse for victims?	What are company resources allocated to content moderation, including the linguistic and cultural contextual expertise of those teams? What is the balance of AI and human moderation in detecting harassment and abuse, and what data is used to train those systems?	

Examples of Current Transparency Initiatives

Issue Area	Policies	Processes	Outcomes
Content & moderation	Tech Against Terrorism mentoring for smaller tech platforms' terms of service	FSM Social Science One	Legally mandated NetzDG transparency reports
Appeals & redress		Facebook Oversight Board FSM as regulated self-regulator as part of the NetzDG	Public reports on decisions made as part of the FSM appeals process
Advertising			Facebook and Google political ad archives
Platform architecture & design	TikTok algorithmic criteria		

Potential Future Transparency Requirements

Issue Area	Policies	Processes	Outcomes
Content & moderation		Regulated reporting on the no. of content moderators, languages, expertise etc. OECD Voluntary Transparency Protocol	
Appeals & redress	Human rights impact assessments		Transparency reporting on volume of and decisions regarding user appeals
Advertising	Regulated advertising transparency requirements for all ads		
Platform architecture & design		Regulator interviews with data scientists adjusting algorithmic design	Algorithmic audit – control tests Data portability

ISD | Powering solutions to extremism and polarisation

Digital Policy Lab
Discussion Paper

National & International Models for Online Regulation

About This Paper

This discussion paper is a snapshot of the evolution of the international policy debate around approaches to countering disinformation, hate speech, extremism & terrorism online. It summarises the policy initiatives presented during the second Digital Policy Lab event on 10–11th December 2020 and further presents a non-exhaustive overview of additional non-regulatory and regulatory policy initiatives of key Five Eyes and EU countries to date, providing a framework for discussion on the benefits, pitfalls and constraints of existing efforts in this space.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of Digital Policy Lab participants or governments.

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Introduction

As communication technology continues to develop with incredible speed and sophistication, extremists across the ideological spectrum have understood how to manipulate online products, media systems and platforms to deceive audiences, distort the available flow of information and conduct illegal activities. Foreign, domestic and transnational groups are deploying tactics to promote social and political polarisation, muddy the availability of accurate, transparent information, subvert democratic processes and spread exclusionary and extreme political agendas. This set of new practices in communication and political mobilisation is emerging far quicker than the framework for understanding or responding to it.

During elections, the norms that have guided what is legitimate and illegitimate political campaigning have been thrown into question by the emergence of this host of new technologies. The ecosystem of social media and communications tools has provided state and non-state actors alike with new levers of influence online and new forms of anonymity with which to cover their tracks. Foreign states and transnational extremist networks have launched campaigns spreading false information about political candidates, parties, governments, activists and minority communities in the US, Germany, Sweden, Italy, France, Mexico, Brazil – to name just a handful of recent examples. Researchers, governments and technology companies have begun to compile evidence detailing the concerted and multi-faceted attempts made to dupe audiences, often with the aim of promoting intolerance, outrage and even violence. But there is much to be done to start responding effectively and proportionally to these complex challenges, which so often connect the threats of hate speech, extremism and disinformation.

Beyond elections, efforts to address some of the most pressing transnational challenges of our time, including climate change, migration, and public health, have come under constant attack by bad actors engaging in manipulative and often illegal activities online. We are witnessing a growing intersectionality in such malign information operations. For example, in the current context of COVID-19, climate science deniers have often been at the forefront of online disinformation efforts to downplay the severity of the pandemic. These forces are not new, but a set of hybrid threats has been hyper-charged by digital technology. Never before have a committed few been able to reach so many so fast with their ideas, boosted by the algorithmic amplification of sensationalist and extreme messaging on social media.

Governments, technology companies, electoral commissions and civil society groups are all trying to deal with the implications of these new developments. Regulation has not yet caught up with the rapid advance of advertising, amplification and audience segmentation technologies. Definitions remain fluid and contested. Questions over intention and outcome still matter deeply in delineating malign uses of technology from legitimate ones, yet these are often the hardest and most politically charged differences to adjudicate. Democracies have grappled with the question of how digital policymaking can effectively safeguard democratic processes, social cohesion and public safety, while also protecting the rights of internet users. Managing speech and information in a liberal society is a painstaking exercise in slow-moving regulation, care and caution, and timidity and patience is easily exploited in the fast-moving world of technology. Historically, the internet has been a powerful tool in projecting liberal values. Equality of information, access to media, and freedom of expression are written into the protocols and infrastructure on which the Internet is built. Protecting those opportunities while preventing their abuse is no small challenge.

Approaches to Digital Policy

Over two decades have passed since the Communications Decency Act first became law in the US and the E-Commerce Directive was adopted by the European Union, providing the underlying liability regimes on which significant parts of the public internet would come to be built. At the heart of these policies was the rationale that freedom of speech and a thriving internet economy could only be guaranteed if online service providers (intermediaries) were not held liable for the user-generated content hosted on their platforms. Since then, the rising threats of disinformation, hate speech, extremism and covert interference on social media have demonstrated some of the limits of this approach.

Initially, numerous **self-regulatory or co-regulatory initiatives** emerged in these areas, with attempts to encourage or cooperate with online platforms to tackle both illegal activity such as terrorism or child abuse, and 'legal harms' such as disinformation or content promoting self-harm. Alongside this, a variety of other approaches to the challenges of online hate speech, extremism, terrorism and disinformation have been implemented, including counter-communications, digital and media literacy, and public awareness campaigns.

Despite the improvements in certain areas through informal, voluntary or industry-led approaches, many governments have still felt compelled to re-open debates on regulating the digital sphere to address these challenges more effectively. In general, this emerging trend towards new online regulation can be dived into two broad categories:

- **Content-based approaches**, often targeting a specific online harm such as hate speech or electoral disinformation, and focusing on the effective and timely removal of that content where appropriate.

- **Systemic approaches**, whereby online platforms must demonstrate that their policies, processes and systems are designed and implemented with respect to the potential negative outcomes that could occur, across a range of possible harms.

Additionally, while outside the scope of this paper, the EU and an increasing number of countries are looking to competition and privacy law to regulate dominant online service providers, with some of the measures having potential beneficial impacts in terms of reducing online harms.

Terrorism & Violent Extremism

The challenge of responding to terrorist and extremist activity on online platforms has been a major concern for governments for well over a decade. However, the issue began to receive major attention from 2014, with the wide-scale proliferation of ISIS propaganda material and recruitment networks across mainstream social media, which accompanied the terrorist group's advance across Iraq and Syria. Increasingly since the terrorist attack in Christchurch, there has also been a focus on extreme far-right content and networks online. Governments and platforms alike have struggled with the scale of content, the resilience of online networks, and the complex jurisdictional considerations and human rights implications of terrorist content and account removal.

In terms of **self-regulation or co-regulation**, the **European Commission** launched the EU Internet Forum in December 2015. The Forum brought together EU Interior Ministers, high-level representatives of major internet companies, Europol, the EU Counter Terrorism Coordinator, and the European Parliament, with the goal of establishing a joint, voluntary approach based on a public-private partnership to detect and address terrorist material online.

This was followed in 2017 by the establishment of the Global Internet Forum to Counter Terrorism (GIFCT), a cross-industry effort to prevent terrorists and violent extremists from exploiting digital platforms, encouraged by the **EU** Internet Forum and the **UK** Home Office. This included the establishment of a cross-platform hash-sharing database of violent terrorist imagery and propaganda, which now contains over 200,000 unique digital 'fingerprints'. In 2019, the Content Incident Protocol (CIP) was developed by the GIFCT to help enable platforms to respond to terrorist events in real time, to assist coordination and prevent the sharing of emergent terrorist content. This protocol was first activated on 9 October 2019 following the terror attack in Halle, Germany after the attacker livestreamed the shooting on the streaming service Twitch, which was not a GIFCT member at the time.

A number of other notable international initiatives have emerged that seek to curb terrorist content online, including the Christchurch Call, initiated by **New Zealand** Prime Minister Jacinda Ardern and **French** President Emmanuel Macron following the 15 March 2019 terrorist attack in Christchurch. The call provides an action plan for collective responses by tech companies, governments and international organisations to eliminate online violent extremist content. It also represents the first major international initiative to put far-right terrorist content on the international agenda, widening a conversation previously dominated by discussions around Islamist terrorism.

In parallel, we have seen the development of additional international dialogues on countering terrorist use of the internet such as the Aqaba Process, launched by **Jordan's King Abdullah II**, and the Global Counterterrorism Forum's Zurich-London Recommendations on Preventing and Countering Violent Extremism and Terrorism Online. Recognising a capacity gap among smaller digital platforms, the **United Nations** Counter Terrorism Executive Directorate's (UN CTED) Tech Against Terrorism initiative, and the **OECD**'s project on voluntary transparency reporting, work with private sector companies to tackle terrorist abuse of the internet whilst respecting human rights.

Other non-regulatory approaches such as **communications-focused efforts** to counter terrorist and extremist propaganda have also been implemented by national governments and at the international level, often in collaboration with the private sector. Initiatives such as the Global Engagement Center at the **US** State Department and Counter Daesh Communication Cell led by the **UK** Foreign, Commonwealth and Development Office with 82 international partners, have developed strategic communications and counter-speech efforts, aimed at addressing the 'demand' side of the equation of online terrorist content. Efforts have adopted both upstream preventative communications approaches aimed at building resilience to violent extremist or terrorist narratives among a broader audience, and downstream strategies aimed at directly rebutting, refuting or countering the narratives of violent extremist or terrorist groups.

In the regulatory space, the majority of efforts to counter violent extremist or terrorist material to date have relied on regulating user content through a notice-and-takedown model, borrowed from established copyright law and specifically the **US**' Digital Millennium Act (DMCA) of 1996. The rationale for these **content-based approaches** is that 'what is illegal offline is also illegal online', and social media companies must be held responsible for the content on their platforms once notified or alerted to it.

In the **UK**, The Counter Terrorism Internet Referral Unit (CTIRU) was set up in 2010 to seek the removal of unlawful terrorist material based on existing legislation. Content that incites or glorifies terrorist acts can be removed under Section 3 of the Terrorism Act 2006 in the UK. Run by the Metropolitan Police, CTIRU compiles a list of URLs for material hosted outside of the UK which are blocked on networks of the public estate and refers content to internet companies for removal. During an average week, the CTIRU remove over 1,000 pieces of content that breach terrorism legislation, the UK government stated in 2015.

Recognising the need for international coordination, **Europol**'s Internet Referral Unit was established in 2015 to flag terrorist and violent extremist online content, share it with relevant government partners, and refer this to companies hosting the content for removal. While the EU IRU has no legal power to compel companies to take down content, parallel referral units and mechanisms have been developed in France, The Netherlands, Belgium, Germany, and Italy.

In September 2018, the **European Commission** presented a proposal that would force social media companies to remove terrorist content within 1 hour of receiving notice from national authorities. If adopted in its current form, the regulation would go beyond a notice-and-takedown model by enshrining a proactive, **systemic approach** through a 'duty of care' obligation, by which companies must ensure their platforms are not used to disseminate terrorist content in the first place. Similarly, a specific Code of Practice for combatting terrorism online will be included in the UK's upcoming Online Harms legislation, due to be introduced to Parliament in 2021 (see below).

Hate Speech

Over the past five years, there have been a variety of government-led initiatives to commit social media companies to fight hate speech on their platforms through self-regulation. In June 2016, the European Commission launched a Code of Conduct on countering illegal hate speech online and invited social media platforms to become signatories. Participating companies voluntarily committed to improving their response time as regards illegal hate speech, as well as bolstering staff training and collaboration with civil society. The EU Code follows the voluntary commitments already made by industry representatives in December 2015 as part of the German task force on online hate led by Heiko Maas (then Minister for Justice and Consumer Protection).

In contrast to the European Commission's assessment of its own Code of Conduct, Jugenschutz.net, the body charged by the **German** task force to monitor industry compliance, found severe shortcomings in the removal of hate speech under these self-regulatory approaches. In response, the German Ministry of Justice adopted a **content-based regulatory approach**, and presented a draft law to combat online hate in March 2017. The draft law was criticised by, among others, the UN Special Rapporteur for Freedom of Expression, who raised concerns about potential 'overblocking' of user content by social media companies. The draft was signed into law with minor changes in June 2017.

Since January 2018, **Germany**'s Network Enforcement Act (NetzDG) has required social media companies with more than 2 million German users to take timely action on illegal content shared on its platforms after it has been flagged by users. Generally, social media companies have seven days to take action on flagged content. This time is reduced to 24 hours when the content flagged is 'manifestly illegal', and failure to comply can lead to fines of up to € 50 million. The law also requires companies to produce a biannual audit on efforts to reduce hate speech on their platform, including figures on reports received from users and the respective actions taken. Users targeted by hate speech may also request data from the platform on their specific case and refer it to the courts through which the perpetrator may be identified. Social media companies can also delegate content moderation to a 'regulated self-regulator', the FSM.

In July 2019, Facebook was fined € 2 million by the German Ministry of Justice for failing to accurately disclose flagging and take-down statistics in its biannual reports. In June 2020, an amendment passed parliament that requires platforms to report certain types of flagged criminal content directly to law enforcement authorities. In September 2020, an independent judicial review requested by the Ministry of Justice found the law was broadly effective in reducing hate speech on platforms, but recommended more user-friendly flagging processes, improving transparency requirements for social media companies and strengthening the ability of users to challenge content decisions made against them.

France has made similar content-based efforts through a proposed law 'on countering online hatred', dubbed the 'Avia law' after its main sponsor in the French National Assembly, Laetitia Avia. In addition to the 24-hour deadline for removing content, similar to the NetzDG, the law stated a 1-hour deadline for the removal of terrorist content or child abuse material, and fines of up to € 1.25 million for failures to act. Unlike the NetzDG, the implementation of the law would have been overseen by an independent regulator, the French High Audiovisual Council (CSA). In 2020, however, the majority of the law was struck down by the French Constitutional Council for infringing on freedom of speech, with concerns raised around the possibility of disproportionality and the potential for 'overblocking' of content.

Most recently, the **Austrian** government, led by the Chancellery and the Ministry of Justice, has proposed a similar law called the 'Communication Platforms Act' in September 2020, targeting platforms with more than 100,000 users or annual revenue exceeding € 500,000. The proposal foresees a dedicated complaints procedure for users to have their content reinstated in cases where removal by the platform is deemed unwarranted (e.g. by the supervisory body KommAustria). Alongside potential fines up to € 10 million, the Act references indirect financial pressure which could be levied in cases of non-compliance, such as blocking the payment of ad revenue to platforms. The Austrian government has no plans to introduce reporting requirements to law enforcement, unlike the recent German amendment.

Disinformation & Foreign Interference

In contrast to efforts to tackle hate speech, violent extremist or terrorist material, initiatives around disinformation have mainly focused on electoral contexts and entailed heightening transparency requirements for political campaigns that use social media, search engines and other open platforms. More recently, the COVID-19 pandemic and spread of health misinformation online has expanded the scope of (self-)regulatory efforts, but has so far mainly resulted in further policy changes by social media companies – with mixed success. Regulating or legislating against malicious uses of information is a challenge fraught with potential pitfalls, unintended consequences and potential risks to free speech and expression. Yet the nature of the threat necessitates a rapid and comprehensive response from policymakers in order to protect democratic processes and societies targeted with dangerous intentional falsehoods that have the potential for harm to individuals and societies.

Self-regulatory approaches were among the first to be introduced to combat disinformation. In September 2018, the **European Commission** published its Code of Practice on Disinformation, which provides an opt-in framework to combat the spread of false or misleading content online. The Code calls for greater ad transparency, active measures against fake accounts, improving the visibility of trustworthy sources and working with the research community to provide better data access, and has since been signed by Facebook, Google, Twitter, Mozilla, Microsoft, TikTok and advertising industry representatives. The implementation of the Code is monitored regularly both by the signatories and the Commission, although progress to date has been variable according to official reporting, in line with ISD's own evaluation that found numerous issues with compliance and efficacy. Despite some improved public scrutiny, the Code has revealed the limits of self-regulation, the potential value added from targets and key performance indicators, and a continued disconnect between platforms, researchers and civil society that continues to hinder efforts to challenge disinformation comprehensively.

Alongside self-regulatory approaches, **digital and media literacy** are perhaps the most frequent non-regulatory measures used to improve resilience to disinformation among the public. Most countries have launched or proposed efforts to this effect, although provision is still highly variable within and across borders. At a regional level, the newly-formed European Digital Media Observatory is tasked with streamlining approaches to counter disinformation, from both a research and education perspective. This includes coordinating academic efforts, improving tools for data analysis, training researchers and journalists, scaling fact-checking efforts, and building a database of materials and best practice for educators.

Beyond voluntary initiatives, various countries have amended their electoral laws to combat online disinformation through **content-based approaches**. **Canada** passed the Election Modernization Act in December 2018, which means entities intentionally selling ad space to 'foreigners who unduly influence an elector' can be fined or even face imprisonment. The Act further requires social media companies to maintain a digital registry of all ads relating to federal elections, including the name of the individual who authorised the advert. In addition, it is an offense 'to make false statements about a candidate for the purpose of influencing the outcome of an election'.

Australia adopted similar transparency rules in March 2018, requiring detailed authorisation statements for all paid advertising distributed on social media during an election. The Foreign Influence Transparency Scheme Act introduced in December 2018 further expands disclosure requirements, including people or entities who produce and disseminate information publicly on behalf of a foreign principal. The bipartisan Honest Ads Act introduced to the **US** Senate in October 2017 broadly mirrors the principles and measures implemented in Canada and Australia, but has not yet been adopted.

In December 2018, **France's law** against the 'manipulation of information' introduced some of the most sweeping transparency requirements seen to date. It requires social media companies to provide details about the funder and money spent on political ads and publish statistics on how platform algorithms promote content related to 'a debate of national interest'; this includes the role of personal data in targeted content dissemination. In addition, companies must provide users with the ability to flag content they deem to be misleading or fake. This requirement builds on the French Electoral Code, which explicitly forbids the distribution of fake news and is punishable by fines or, in some cases, imprisonment. The 2018 law further empowers judges to order any 'proportional and necessary measure' to halt fake or misleading information from spreading online. Claims can be brought by 'any person interested in acting' (personne ayant intérêt à agir), and decisions must be rendered within 48 hours.

Towards Systemic Approaches[1]

Many of the online harms and activities coordinated by extremist groups or foreign states that regulations seek to tackle cross existing legal thresholds. Some are newly regulated, such as the creation of disinformation identified during election periods in France. Others continue to promote vigorous legislative debate, such as the proscription of bots used to promote political campaigns in Ireland. But most still sit in a grey zone of acceptability, straddling technology company terms of service and national laws which yet have to catch up with the evolving threat. Furthermore, the borderless internet does not render attribution an easy task. The line between state actors and non-state networks is increasingly difficult to distinguish, as is the ability to trace the perpetrators of illegal hate speech, extremist content or disinformation.

These issues have led some legislators to look beyond content-based regulation and adopt a cross-harms perspective to online regulation. The aim of these more 'systemic' approaches is to develop an oversight framework that can be used to tackle a plethora of online harms, ranging from hate, extremism and terrorism to child safety, cyber-bullying and disinformation.

In the **UK**, the Online Harms White Paper (published April 2019) laid out various legislative measures that could be used to regulate a broad set of issues, with the goal of making the UK 'the safest place to be online'. At the heart of the proposed framework is a statutory duty of care, overseen by a central regulatory body such as Ofcom. This would enshrine a legal obligation for online service providers to 'keep their users safe from harm'. In stark contrast to approaches that seek to enforce existing laws online such as the German NetzDG, the UK regulation would encompass content that is not illegal but still deemed harmful (so-called 'legal harms'). Much like the European Commission's proposal surrounding terrorist material, this approach would force companies to adopt (and evidence) proactive measures that prevent harmful content appearing or gaining exposure in the first place, rather than merely responding to third-party notice-and-takedown requests for specific pieces of content.

In the **European Union**, a revision of the E-commerce directive is underway in the form of the Digital Services Act (DSA). In its initial impact assessment, the European Commission identified a broad set of issues that the governance of digital services in the EU must address, including societal harms, illegal activity and insufficient protection of fundamental rights. Notably, ineffective supervision and insufficient administrative cooperation were also identified as key shortcomings of the current regime, indicating plans to move beyond a notice-and takedown regime towards more systemic oversight of social media platforms by European regulators. In addition, the European Democracy Action Plan highlights the need for a combination of regulatory and non-regulatory initiatives to protect elections, safeguard media pluralism and combat disinformation on a European level – all of which require increased cooperation and, in some cases, co-regulation, across public and private sectors.

Even in the case of the content-focused **German** NetzDG, a proposal to revise the law include an oversight mandate for the Federal Office of Justice. All of these developments indicate that, as the 20th anniversary of the E-Commerce Directive passes, the new generation of regulatory and non-regulatory initiatives to combat hate, extremism and terrorism will increasingly throw the internal processes of social media companies into the limelight. Alongside continued debates about what constitutes harmful online content, the emphasis will be on ensuring regulators have the proper tools at their disposal to fulfil their oversight function while safeguarding their operational and functional independence – a marked evolution of digital regulation in a space that has to date been dominated by a (often voluntary) notice-and-takedown model.

[1]Please also consult our separate briefing paper with a specific focus on the DSA and UK Online Safety Bill.

ISD | Powering solutions to extremism and polarisation

Digital Policy Lab
Policy Summary

EU Digital Services Act & UK Online Safety Bill

About This Paper

This paper provides a non-exhaustive summary of the digital policy proposals of the EU and the UK, namely the Digital Services Act and the Online Safety Bill published in December 2020. It accompanies the Digital Policy Lab session taking place on 27th January 2021 and focuses on key elements of the two proposals relevant to combatting terrorism, extremism, hate and disinformation online.

The views expressed in this document are those of the authors and do not necessarily reflect the views of Digital Policy Lab participants or governments.

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

The EU's Digital Services Act (DSA)

On 15th December 2020 the European Commission presented a new set of rules for all digital services, including social media, online marketplaces, and other online platforms that operate in the European Union: the Digital Services Act and the Digital Markets Act.

The Digital Services Act (DSA)[1] regulates the obligations of digital services that act as intermediaries in their role of connecting consumers with goods, services, and content. They should "create a safer online experience for citizens to freely express their ideas, communicate and shop online by reducing their exposure to illegal activities and dangerous goods and ensuring the protection of fundamental rights".[2] As such it updates the e-Commerce Directive, which has been in place since 2000.

According to the European Commission, the DSA preserves "a balanced approach to the liability of intermediaries, and establishes effective measures for tackling illegal content" and "societal risks online".[3] The Commission states that these new rules are "an important step in defending European values in the online space", and it aims at "setting a benchmark for a regulatory approach to online intermediaries also at the global level".

The DSA was presented together with the Digital Markets Act (DMA). This Act imposes new rules on so-called 'gatekeeper' companies, which are platforms that serve as an important gateway for other businesses to reach their customers. These companies control at least one so-called "core platform service" (such as search engines, social networking services, certain messaging services, operating systems and online intermediation services), and have a lasting, large user base in multiple countries in the EU. Under the Digital Markets Act, companies identified as gatekeepers will need to proactively implement certain behaviour, and will have to refrain from engaging in unfair behaviour. The DMA aims to prevent those gatekeepers from imposing unfair conditions on businesses and consumers and to ensure the openness of important digital services. Examples include prohibiting businesses from accessing their own data when operating on these platforms, or situations where users are locked into a particular service and have limited options for switching to alternative services (denial of interoperability).

The proposals from the European Commission provide the starting point for both the European Parliament and Member States to adopt legislation at European Union level. As co-legislators they will first amend the proposals along their preferences before agreeing on a compromise text. This procedure is expected to last for up to three years. The following sections will present a summary of the most important features of the Digital Services Act.

New Horizontal Rules for All Categories of Content vs. Sector-Specific Rules

The DSA proposal updates the horizontal rules that define the responsibilities and obligations of providers of digital services, and online platforms in particular, in the European Union. These rules apply in the EU without discrimination, including to those online intermediaries established outside of the European Union that offer their services in the EU.

As such, it introduces a horizontal framework for all categories of content, products, services and activities on intermediary services.[4] For the purposes of this briefing note it is useful to clarify how the proposal deals with three different, often debated categories of content: harmful content, illegal content and manifestly illegal content.

The proposal recognises how the growth of certain services has "increased their role in the intermediation and spread of unlawful or otherwise harmful information and activities" (§5). However, the proposal does not define **harmful information and activities**, arguing that there is

> *"a general agreement among stakeholders that 'harmful' (yet not, or at least not necessarily, illegal) content should not be defined in the Digital Services Act and should not be subject to removal obligations, as this is a delicate area with severe implications for the protection of freedom of expression" (p.9).*

Instead, the Commission creates **due diligence obligations** for platforms' content moderation activities, which includes "activities aimed at identifying and addressing illegal content or information incompatible with their terms and conditions" (Article 2.p). What is seen as harmful content, for example in the UK's Online Safety Bill, is covered mainly by this reference to 'information that is incompatible with the terms and conditions of a platform'. The Commission states in its supporting materials that; "to the extent that it is not illegal, harmful content should not be treated in the same way as illegal content".[5] Rules to remove or encourage removal of content (cf infra) will only apply to illegal content, in full respect of the freedom of expression.

Illegal content is any information related to "illegal content, products, services and activities" as defined by Union law or a national Member State law (Article 2g, §12). Hence, the DSA does not touch upon national or EU laws that specify what constitutes illegal content. The DSA covers information that by itself is illegal, such as hate speech, terrorist content, unlawful discriminatory content, but also information that relates to activities that are illegal, such as the sharing of child sexual abuse material (CSAM), non-consensual sexual images (so-called 'revenge porn') or online stalking or harassment.

Manifestly illegal content is content where it is evident to a layperson, without any substantive analysis, that the content is illegal (§47). This is relevant only in the context of Article 20 (cf. infra, Section 4.3), which states that online platforms can suspend users from their service who frequently provide such content.
This proposed regulation complements existing sector-specific legislation, such as the Audio-Visual Media Services Directive (AVMSD), the Copyright Directive, the Consumer Protection Acquis or the future terrorist content directive[6], which apply as *lex specialis* (§9), meaning that the earlier sector specific obligations will override the DSA if it governs the same topic in less detail. By way of example, specific obligations that were imposed on video sharing platforms like YouTube to counter hate speech in the AVMSD will continue to apply. This Regulation would apply then to those video sharing providers to the extent that the AVSMD does not - or does not fully - address these topics (DSA, p.4).

Updated Framework for the Conditional Exemption from Liability of Providers of Intermediary Services

Chapter II of the DSA includes the conditions under which providers of mere conduit (Article 3), caching (Article 4) and hosting services (Article 5) are exempt from liability for the third-party information they transmit and store. The definitions of these intermediary services have remained the same (see Article 2f). Articles 3 and 4 are copies of Articles 12 and 13 from the e-Commerce Directive. These rules apply to every provider of an intermediary service "irrespective of their place of establishment or residence" (Article 1.3), insofar as they provide services in the EU as evidenced by a "substantial connection" to the EU. A substantial connection could be derived from specific factual criteria, such as the number of users in the EU, or the targeting of activities towards one or more Member States (Article 2d, §7-8).

This proposal now for the first time introduces **'online platforms' as a subcategory of hosting services** (see Article. 2f, §13). These are seen as providers of hosting services that not only store information provided by the recipients of the service at their request, but that also disseminate that information to the public, again at their request (Article 2h). This new category is not important from a liability perspective, but is important for determining the categories of providers to which the new due diligence obligations apply (cf. infra Section 4). Interpersonal communication services like Telegram or WhatsApp are not covered by this definition, and neither are "intermediary services" such as "remote information technology services, transport, accommodation or delivery services" (§6).

The proposal **maintains the liability rules for providers of intermediary services as interpreted in the last two decades by the Court of Justice** (§16). The rule of thumb is still: if a hosting service obtains actual knowledge or awareness of illegal content, it needs to act expeditiously to remove or to disable access to that content (Article 5.1, §22).[7] The exemptions from liability do not apply to providers which play "an active role of such a kind as to give it knowledge of, or control over" information provided by a user (§18).

Another new element of the DSA is the introduction of what is often referred to as a **Good Samaritan clause**. The new Article 6 aims to eliminate existing disincentives towards voluntary, proactive investigations undertaken by providers of intermediary services (or when they comply with national law). It also clarifies that undertaking such investigations aimed at detecting, identifying and removing, or disabling access to, illegal content does not make a platform ineligible for the liability exemptions (Article 6, §25).

Article 7 lays down a **prohibition of general monitoring** or active fact-finding obligations for those providers (Article 7) - similar to Article 15 of the e-Commerce Directive. Finally, this section of the DSA imposes an obligation on providers of intermediary services in respect of orders from national judicial or administrative authorities to act against illegal content (Article 8) and to provide information (Article 9).

Tiered Structure of Due Diligence Obligations for Different Sorts of Intermediary Services

On top of providing a framework for the conditional exemption on the provision of intermediary services in the internal market, the DSA also introduces new due diligence obligations that are adapted to the type and nature of the intermediary service concerned. There are a set of basic due diligence obligations that apply to all providers of intermediary services, which are then complemented by additional obligations for providers of hosting services and online platforms.

For very large online platforms (VLOPs) the proposal sets asymmetric due diligence obligations depending on the nature of their services and their size. The proposal sets up a "supervised risk management approach" (DSA, p.11), in which certain substantive obligations are limited only to VLOPs which due to their reach have acquired a central, systemic role in facilitating the public debate and economic transactions (§53). This approach addresses certain identified problems only where they materialise, while not overburdening providers unconcerned by those problems. This results in the following set of obligations for four different categories of intermediary services:

<table>
<tr><th>Intermediaries</th><th>Hosting Services</th><th>Online Platforms</th><th>VLOPs</th></tr>
<tr><td colspan="4">Transparency reporting
Requirements on terms of service and due account of fundamental rights
Cooperation with national authorities following orders
Points of contact and, where necessary, legal representative</td></tr>
<tr><td rowspan="3"></td><td colspan="3">Notice and action, and information obligations</td></tr>
<tr><td colspan="2"></td><td>Complaint and redress mechanism and out of court settlement
Trusted flaggers
Measures against abusive notices and counter-notices
Vetting credentials of third party suppliers ("KYBC" – "know your business customer")
User-facing transparency of online advertising</td></tr>
<tr><td colspan="2"></td><td>Risk management obligations
External risk auditing and public accountability
Transparency of recommender systems and user choice for access to information
Data sharing with authorities and researchers
Codes of conduct
Crisis response cooperation</td></tr>
</table>

Obligations for All Providers of Intermediary Services

Obligations for all providers of intermediary services are laid out in Section 1 of Chapter 3 and include the following four main obligations:

- The **obligation to establish an operational single point of contact** to facilitate direct communication with Member State authorities, the European Commission, and the European Board for Digital Services ('the board' - cf infra) (Article 10, §36).

- The obligation to designate a **legal representative** in the Union for providers not established in any Member State, but offering their services in the Union (Article 11, §37). The designated legal representative can be held liable for non-compliance with obligations under this Regulation.

- A transparency obligation to **set out in their terms and conditions any restrictions that they may impose on the use of their services**. That information shall include any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review (Article 12.1, §38). Providers must act responsibly in applying and enforcing those restrictions (Article 12.2).

- Providers of intermediary services should **annually report on the content moderation they engage in**, irrespective of whether this is illegal content or contrary to the providers' terms and conditions (Article 13, §39). 'Very large platforms' (cf infra) should report every six months. This includes mainly aggregated data about:

  - The number of take down orders received from member states authorities (13.1a);

  - The number of notices submitted (13.1b);

  - The number of complaints received about its content moderation decisions (Article 13.1.d);

  - Interestingly, it also includes information on "the content moderation engaged in at the providers' own initiative, including the *number and type of measures* taken that **affect the availability, visibility and accessibility of information** provided by the recipients of the service and the recipients ability to provide information, categorised by the type of reason and basis for taking those measures" (Article 13.1.c).

Platforms and very large platforms are subject to extra transparency requirements (see Articles 23 & 33).

Obligations for Hosting Services

Section 2 of Chapter 3 lays down obligations, additional to those under Section 1, that are applicable to all providers of hosting services. This would include **file storage and sharing services, web hosting services, advertising servers and 'paste bins'**, in as far as they qualify as providers of hosting services covered by this Regulation. In particular, that section obliges those service providers to put in place user friendly notice and action mechanisms to allow third parties to flag the presence of alleged illegal content (Article 14, §40, 41). Article 14.2 includes the elements that need to be included in a notice for it to be considered giving rise to actual knowledge.

Furthermore, if such a provider decides to remove or disable access to specific information provided by a recipient of the service, they must provide that recipient with a statement of reasons and the available redress possibilities (Article 15, §42).

Obligations for All Online Platforms

Section 3 of Chapter 3 lays down obligations applicable to all online platforms, additional to those under Sections 1 and 2. This Section does not apply to online platforms that are micro or small enterprises "unless their reach and impact is such that they meet the criteria to qualify as very large online platforms under this regulation" (Article 16, §43).

- They must **provide an internal complaint-handling system** in respect of decisions to remove or disable access taken in relation to alleged illegal content or information incompatible with their terms and conditions (Article 17). This should allow users to easily and effectively contest certain content moderation decisions (§44).

- They must engage with certified **out-of-court dispute settlement bodies** to resolve any dispute with users of their services (Article 18, §44/45).

- They must expedite notices submitted by entities (not individuals) that are granted the status of **trusted flaggers** by the national Digital Services Coordinator (Article 19, §46).

- It sets out the **measures online platforms are to adopt against misuse** (Article 20, §47), respectively for when users are frequently providing manifestly illegal content or by frequently submitting manifestly unfounded notices or complaints. Under certain conditions, online platforms should temporarily suspend "their relevant activities in respect of the person engaged in abusive behaviour" (§47).

- The Section includes an **obligation to inform competent law enforcement authorities** in the event they become aware of any information indication that a person "may have committed, may be committing or is likely to commit a serious criminal offence involving a threat to the life or safety of person" (Article 21, §48).

- The Section obliges online platforms to **assess the reliability of and publish specific information on the traders using their services** where those online platforms allow consumers to conclude distance contracts with those traders (Article 22, §49).

- Those online platforms are also obliged to organise their interface in a way that enables traders to respect Union consumer and product safety law (Article 22, §50).

- **In addition to their transparency obligations in Article 13**, online platforms are also obliged to include in their annual reports data on:

  – The number of disputes submitted to the out-of-court dispute settlement bodies;

  – The number of suspensions imposed pursuant to Article 20;

  – Any use made of automatic means for the purpose of content moderation, including a specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards applied (Article 23, §51).

- The Section includes **additional transparency obligations for online platforms that display advertising** (Article 24, §52). The proposal highlights how online ads can contribute to significant risks since they (1) can include illegal content, (2) provide financial incentives for the publication or amplification of illegal or otherwise harmful content and activities online or (3) can be targeted in a discriminatory way with an impact on the equal treatment and opportunities of citizens. As a result, the DSA imposes extra user-facing transparency measures on those platforms that would allow users to identify for each ad in real time:

  – That the information displayed is an ad;

  – The natural or legal person on whose behalf the ad is displayed;

  – Meaningful information about the main parameters used to target ad recipients, which can include "providing meaningful explanations to the logic used to that end, including when this is based on profiling" (§52);

  The Code of Conduct on online advertisements in Article 36 is supposed to further determine such 'meaningful information' about the 'main parameters' should work (cf Article 36.2.b).

Risk Assessment, Risk Mitigation & Auditing Obligations for VLOPs

Section 4 lays down obligations, additional to those laid down in Sections 1 to 3, for very large online platforms (as defined by Article 25, §53-55) to manage systemic risks. The **operational threshold for service providers in scope of these obligations** includes those online platforms with a significant reach in the Union, currently estimated to be more than 45 million average monthly active recipients in the EU. Where the EU's population changes by a certain percentage, the Commission will adjust the number of recipients considered for the threshold, so that it consistently corresponds to 10% of the Union's total population.

Once a platform reaches this threshold, the systemic risks it poses can have a disproportionately negative impact on our societies given their reach and ability to facilitate public debate and disseminate information online. The Commission argues that the way these platforms design their services "is generally optimised to benefit their often advertising-driven business models and can cause societal concerns. In the absence of effective regulation and enforcement, they can set the rules of the game, without effectively identifying and mitigating the risks and the societal and economic harm they can cause" (§56).

Hence, VLOPs are **obliged to conduct risk assessments** at least once a year on the systemic risks stemming from the functioning and use of their service, as well as by potential misuses, and then take appropriate mitigating measures (Article 26, §57). The proposal identifies the following systemic risks:

- (26a) **the dissemination of illegal content through their services**; such as the dissemination of child sexual abuse material or illegal hate speech, and the conduct of illegal activities, such as the sale of counterfeit products, for example. The Commission highlights the risks of "accounts with a particularly wide reach" that disseminate such content or engage in this type of conduct (§57).

- (26b) **any negative impact of the service on the exercise of fundamental rights** for private and family life, freedom of expression and information, the prohibition of discrimination and the rights of the child, as enshrined in Articles 7, 11, 21 and 24 of the Charter respectively. This is an important category as the recital makes clear that these risks "may arise, for example, in relation to the design of the algorithmic systems used by the very large online platform" or the misuse of their service through the submission of abusive notices or other methods for silencing speech or hampering competition (§57).

- (c) **intentional manipulation of their service**, including by means of inauthentic use or automated exploitation, with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security, having regard to the need to safeguard public order, protect privacy and fight fraudulent and deceptive commercial practices. Such risks may arise, for example, through the creation of fake accounts, the use of bots, and other automated or partially automated behaviours, which may lead to the rapid and widespread dissemination of information that is illegal content or incompatible with an online platforms' terms and conditions.

The Commission states that when VLOPs are conducting risk assessments they need to take into account in particular "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 the potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions" (Article 26.2).

VLOPs should, where appropriate, conduct their risk assessments and design risk mitigation measures with the involvement of representatives of the recipients of the service, groups potentially impacted by their services, independent experts and civil society organisations (§59).

After having identified those risks, VLOPs should deploy "reasonable, proportionate and effective" **means to mitigate those risks** (Article 27.1, §58). Such measures may include:

- Article 27.1a Adapting their terms and conditions, and the design and functioning of their content

moderation processes, algorithmic recommender systems, online interfaces, or other features of their services. This can include, for example, improving the visibility of authoritative information sources.

- Article 27.1b Targeted measures aimed at limiting the display of ads in association with the service they provide, including by discontinuing advertising revenue for specific content.

- Article 27.1.d/e Initiating or enhancing cooperation with trusted flaggers, including via training sessions and exchanges with trusted flagger organisations, and cooperation with other service providers, including by initiating or joining existing codes of conduct or other self- regulatory measures.

The Commission may issue general guidelines on the application of these specific risk mitigating measures.

Given the need to ensure verification by independent experts, VLOPs should be accountable, **through independent auditing**, for their compliance with the obligations laid down by this Regulation and, where relevant, any complementary commitments pursuant to codes of conduct and crisis protocols. To achieve this goal they are also to submit themselves to external and independent audits (Article 28, §60), which - in case the audit opinion is not positive - results in operational recommendations on specific measures to achieve compliance (Article 28.3.f). Platforms should give the auditor access to all relevant data necessary to perform the audit properly. Auditors should also be able to make use of other sources of objective information, including studies by vetted researchers. The audit report is sent to the Digital Services Coordinator of establishment (cf infra) and the EU Board, which scrutinise whether the proposed recommendations have been properly addressed. If this is not the case the Commission may further investigate (Article 51 - cf infra), and ultimately fine or impose other interim measures on the very large platform. The Section includes **a specific obligation where VLOPs use recommender systems** (Article 29, §62) or **display online advertising** on their interface (Article 30, §63).

The Commission recognises that **recommender systems** can play "an important role in the amplification of certain messages, the viral dissemination of information and the stimulation of online behaviour" (§62). Hence, VLOPs should ensure that recipients are appropriately informed, and can easily influence the information recommended to them. They should clearly present the main parameters for such recommender systems in an easy and comprehensible manner to ensure that the recipients understand how information is prioritised for them. They should also ensure that the recipients enjoy alternative options for the main parameters, including options that are not based on profiling.

On **advertising systems**, the Commission recognises that these pose "particular risks" and "require further public and regulatory supervision on account of their scale and ability to target and reach recipients of the service based on their behaviour within and outside that platform's online interface" (§63). Hence, VLOPs should ensure public access - through APIs - to an ad archive, which includes the content of the ads, the natural or legal person on whose behalf the ad is displayed, the period during which the ad was displayed, the main targeting parameters and the total numbers of recipients the ad reached.

Furthermore, the Section sets out the conditions under which VLOPs shall provide **access to data** via online databases or APIs to the Digital Services Coordinator of establishment or the Commission that are necessary to monitor and assess compliance with this regulation (Article 31.1, 31.3, §64). Importantly, such a requirement may include, for example, the data necessary to assess the risks and possible harms brought about by the platforms systems, data on the accuracy, functioning and testing of algorithmic systems for content moderation, recommender systems or advertising systems, or data on processes and outputs of content moderation or internal complaint-handling systems. (§64) Upon a request from one of these two actors, VLOPs shall provide access to data to vetted researchers "for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks" (Article 31.2). In order to be vetted, researchers shall be affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the

risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request (Article 31.4).

Finally, VLOPs must appoint one or more compliance officers to ensure compliance with the obligations laid down in the Regulation (Article 32). Additionally, Article 33 lists the specific transparency obligations for VLOPs that are outlined in Articles 26-28.

Means To Implement & Standardise Due Diligence Obligations

Section 5 describes the processes which the Commission will support and promote to facilitate the effective and consistent application of the obligations in this Regulation that may require implementation through technological means. The Commission want to promote "voluntary industry standards" such as Codes of Conducts, which can cover "certain technical procedures, where the industry can help develop standardised means to comply with this Regulation, such as allowing the submission of notices, including through application programming interfaces, or about the interoperability of advertisement repositories" (§66).

The Commission and the Board shall encourage the drawing-up of **codes of conduct** to contribute to the application of this Regulation (Article 35, §67-69). Importantly, these Codes can include commitments to take specific risk mitigation measures (Article 35.2) which can be assessed on the basis of KPIs (Article 35.3). The Commission can invite platforms and other interested parties to voluntarily participate in the Code of Conduct (Article 35.2), but the incentive to participate is substantial given that "adherence to and compliance with a given code of conduct by a very large online platform may be considered as an appropriate risk mitigating measure" (§68). This is what is meant with the "co-regulatory backstop". Codes of Conduct are voluntary, but not participating in them increases the risk of non-compliance with the DSA.

The DSA identifies a number of areas of consideration for such codes of conduct, in particular (1) risk mitigation measures concerning specific types of illegal content, and (2) systemic risks for society and democracy, such as disinformation or manipulative and abusive activities. This includes coordinated operations aimed at amplifying information, such as the use of bots or fake accounts, sometimes with a purpose of obtaining economic gain, which are particularly harmful for vulnerable recipients of the service, such as children (§68).

The Commission lists specific future **code of conduct for online advertising**. This code would go beyond the mandatory ad archives in Article 30 and the user-facing transparency tools about ads in Article 24. The goal of this Code is aimed at bringing different actors together with civil society organisations or relevant authorities to provide more transparency about the "transmission of the relevant information" in the ad tech value chain, from publishers to advertisers (Article 36, rec 70) in order to ensure a "competitive, transparent and fair environment in online advertising" (Article 36.2).

There is also a provision on **crisis protocols to address extraordinary circumstances affecting public security or public health** (Article 37, §71). This includes "any unforeseeable event, such as earthquakes, hurricanes, pandemics and other serious cross-border threats to public health, war and acts of terrorism, where, for example, online platforms may be misused for the rapid spread of illegal content or disinformation or where the need arises for rapid dissemination of reliable information" (§71). VLOPs should be encouraged to establish and apply specific crisis protocols. Such protocols should be activated only for a limited period of time and the measures adopted should be limited to what is strictly necessary to address the extraordinary circumstance, including, for instance by prioritising "prominent information on the crisis situation provided by Member States Authorities or at Union level" (Article 37.2.a).

Supervision & Enforcement

The Regulation works on the basis of three rules of thumb:

- Ensuring adequate oversight and enforcement should in principle be attributed to the Member States.
- The Member State in which the main establishment of the provider of the intermediary services is located shall have the jurisdiction over the due diligence obligations for platforms, which would typically be an Irish regulator for VLOPs.
- Where systemic risks emerge across the Union posed by VLOPs, the Regulation provides for supervision and enforcement at Union level - be it by the European Board for Digital Services or the European Commission.

Digital Services Coordinators (DSCs)	**European Board for Digital Services (Board)**	**European Commission (EC)**
• Independent authorities • Direct supervision and enforcement (by default) • Coordination with the other national competent authorities • Coordination and cooperation at EU level with Board, EC and other DSCs	• Ad-hoc independent advisory group • Composed of DSCs • Chaired by EC, no vote • Advising DSCs and EC, recommending actions • No binding acts, but EC needs to take them into account • Cooperation with other EU bodies, agencies, offices on related matters	• Direct enforcement powers vis a vis very large online platforms for: – Specific obligations for VLOPs (after DSC supervision) – All other obligations (if DSC failed to act) • Administrative support to the Board • Advising on cross-border disputes • Intervening upon DSC request

National-Level Oversight

Every Member State should appoint "at least one" existing or new national authority with the task to apply and enforce this regulation, especially on the liability regime for intermediary services, but specific regulators can be in charge for specific supervisory or enforcement tasks (such as media regulators or consumer protection authorities) (§72). Only one authority can be the 'Digital Services Coordinator', which acts as a single point of contact.

The Digital Services Coordinators have **investigative powers**, including the power to require providers to hand over any information relating to a potential infringement of the Regulation (Article 41a), the power to carry out on-site inspections in order to seize such information (Article 41b) and the power to impel testimony from any company staff member in relation to that information (Article 41c).

They also have **enforcement powers**, including to order the cessation of the infringement, issue fines and adopt interim measures. A failure to comply with the obligations in the DSA can result in a fine of up to 6% of the annual income or turnover of the provider, whereas penalties for supplying incorrect, incomplete or misleading information can result in a fine of up to 1% of the annual income or turnover of the provider (Article 42.3). If the infringement still persists after having exhausted all these powers, and entails serious criminal offences, the Digital Services Coordinator can ultimately request a judicial authority to order the temporary restriction of access to users for at least four weeks (Article 41.3b).

Importantly, the Member States in which the main establishment of the provider of the intermediary services is located shall have **jurisdiction** over the due diligence obligations for that platform, which would typically be an Irish regulator (Article 40.1). However, other Digital Services Coordinators from other jurisdictions - or the Board - can request the Digital Service Coordinator to take the necessary investigatory and enforcement measures (Article 45). Individuals or representative organisations should be able to lodge any complaint related to compliance with this Regulation with the Digital Services Coordinator in the territory where they received the service (Article 43, §81), and where necessary the DSC will refer the complaint to the DSC of establishment.

Supranational Oversight

The DSA would establish a European Board for Digital Services, an independent advisory group at EU level which supports the Commission and helps coordinate the actions of the DSC, including by suggesting appropriate investigation and enforcement measures. The Board would also contribute to drafting codes of conduct. It would give non-binding opinions to the DSCs or other competent national authorities. It would consist of all the Digital Services Coordinators (Article 47, §88-90).

Where systemic risks emerge across the Union posed by VLOPs, the proposed Regulation provides for supervision and enforcement at Union level. Section 3 concerns the supervision, investigation, enforcement and monitoring of VLOPs. It provides for enhanced supervision in the event of platforms infringing the provisions of Chapter III, Section 4 (Article 50, §94).

Once an infringement has been identified, for instance pursuant to individual or joint investigations, auditing or complaints, the Digital Services Coordinator of establishment should monitor any measure taken by the VLOP as set out in its action plan (Article 50.1/50.2). If it has concerns that those measures might not be effective, it can request another audit of those measures (Article 50.3/95). If that infringement hasn't been addressed by the VLOP in the view of the Digital Services Coordinator, the Commission may further investigate (Article 51).

The DSA also provides the possibility for the Commission to intervene vis à vis VLOPs on its own initiative or after a request from the Board (Article 51, §96). In these cases the Commission can carry out investigations, compelling VLOPs to provide any relevant document, data and information necessary for that investigation, including explanations relating to databases and algorithms (Article 52, §99), interviews (Article 53) and on-site inspections (Article 54). During on-site inspections the Commission and auditors (or experts appointed by it) may require the VLOP concerned to provide explanations on its organisation, functioning, IT system, algorithms, data-handling and business conduct (Article 54.3).

The Commission can adopt interim measures (Article 55), and make commitments made by VLOPs binding (Article 56), as well as monitor their compliance with the Regulation (Article 57). In case of non-compliance, the Commission can adopt non-compliance decisions (Article 58), as well as fines (Article 59) and periodic penalty payments (Article 60).

The UK's Online Harms White Paper Consultation Response

The UK Government's consultative stage on proposals to 'make the UK the safest place in the world to go online, and the best place to grow and start a digital business' (1.0)[8] are now complete[9]. The Online Harm proposals were first set out in the Online Harms White Paper published in April 2019, with an interim consultation report published in February 2020, and the full consultation report published in December 2020. The new regulatory framework will continue to progress, slightly rebadged, via the forthcoming Online Safety Bill expected in 2021. The core proposal is for a *new statutory duty to care, enforced by an independent regulatory body, now confirmed to be* Ofcom.

Companies in Scope

The framework applies to companies who either 'host **user-generated content** which can be accessed by users in the UK' or 'facilitate public or private online interaction between service users' with at least one party who is in the UK (1.1). **Search engines** have been clarified to be within scope (1.3). The UK Government has emphasised they will be applying a risk-based approach, focusing regulatory activity on companies whose services pose the biggest risk of harm (20). Activities intended to be excluded from scope include ISPs, hosting providers, app stores as well as business-to-business services (1.2).

The consultation response also states that journalistic content will have particular protections, including exceptions for news media's own websites and "robust protections for journalistic content shared on in-scope services". The stated reason is that media companies have 'raised concerns that regulation may result in increased takedowns of journalistic content' (1.11). No detail has yet been provided on how journalistic content is to be defined, nor how those protections for content shared on in-scope services will work in practice.

While all companies in scope of the proposals will owe a statutory duty of care, only a small number of high reach, high risk companies (2.16) will be designated '**Category 1**' with additional obligations. There will be a three step process for designation:

1. Primary legislation will set out high level factors (2.16) including size of audience and functionality offered;

2. Government will determine thresholds for these factors, following advice from Ofcom;

3. Assessment from Ofcom against these factors and thresholds. (2.18).

While no specific companies were named in the consultation response, media briefings suggested that it was likely Category 1 companies included: 'Facebook, TikTok, Instagram and Twitter' among others such as YouTube. The vast majority of services will be 'Category 2' companies with no additional obligations.

Duty of Care

The aim of the duty of care will be to improve safety for users of online services, primarily by taking action against content or activity that '**cause[s] significant physical or psychological harm to individuals**' (2.7). Companies will 'complete an assessment of the risks associated with their services and take **reasonable steps to reduce the risks of harms** they have identified occurring'. Steps will include 'user tools, content moderation and recommendation procedures' (2.9). The framework will **apply to both public and private** communication channels and services, including for example messaging apps (29).

Ofcom will issue statutory codes of practice on how companies can fulfil the duty of care (2.48), following consultation with stakeholders (2.50). Companies can take alternative measures to those in the codes, so long as they can demonstrate they are equivalent or exceed the standards outlined in the codes (2.48).

Companies will have specific legal duties to implement effective **reporting and redress** mechanisms (2.12), but government will not mandate specific forms of redress (2.13), and there will not be new avenues for individuals to sue companies (2.12). Ofcom will additionally publish codes of practice on redress mechanisms (2.12).

Online Harms

Legislation will set out a definition of the harmful content and activity that 'cause[s] significant physical or psychological harm to individuals' (2.7) in scope of the regime to help provide legal certainty (2.1). Harms which are expressly excluded include those related to intellectual property, data protection, fraud, consumer protection and cyber security breaches (2.4) on the basis they are covered by other regulatory regimes.

Priority categories of harmful content and activity which will be set out in secondary legislation include: i) priority criminal offences e.g. Child Sexual Exploitation and Abuse (CSEA), terrorism, hate crime; ii) content that is harmful to children e.g. pornography, violent content; iii) 'priority categories of harmful content and activity that is legal when accessed by adults, but which may be harmful to them' e.g. abuse, eating disorders, self-harm (2.3 & 2.19). Existing legal responsibilities regarding illegal content and activity will remain in place (2.23). These priority categories cover both illegal, as well as legal-but-harmful content and activities.

Voluntary interim codes of practice[10] dealing with two priority categories, CSEA and terrorism have been published alongside the consultation response (2.51). The CSEA code builds on the Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse (2.53). Ofcom will have powers to require the use of '**automated technology that is highly accurate**' (2.59) to identify CSEA content or activity on platforms, where there are 'no alternative, less intrusive approaches [...] available' (2.62). Ofcom will be required to report annually to the Home Secretary on the use of the power including on effectiveness and accuracy (2.61). Ofcom will also be empowered to mandate automated technology to identify, flag, block or remove (2.42) illegal terrorism content and activity where it is effective, proportionate and necessary (2.70).

All companies will be required to **assess the likelihood of children accessing their services**. If they determine reasonable likelihood, they will be required to provide additional protections for children using them. This is in line with the approach taken by the Information Commissioner's Age Appropriate Design Code which provides standards for protecting children's personal data.

Where **disinformation and misinformation** causes 'significant physical or psychological harm to individuals', it is within scope of the framework. Where mis- or disinformation is 'unlikely to cause this type of harm it will not fall in scope of regulation' (2.81). Decisions relating to 'political opinions or campaigning, shared by domestic actors within the law' are intended to be out of scope (2.81). Ofcom will be required to establish 'an expert working group on disinformation and misinformation [..] to build consensus and technical knowledge on how to tackle disinformation and misinformation' (2.85).

Regulatory Powers & Enforcement

Ofcom will have a number of functions including setting out what companies need to do to fulfil the duty of care, requiring companies to have effective redress mechanisms, establishing **user advocacy** mechanisms (4.39), providing a **super-complaints** function (4.37), and promoting online safety and innovation (Box 16).

Transparency reports (4.15) may be required by Category 1 companies setting out what they are doing to meet the duty of care, and may be extended to Category 2 companies. These are likely to include information about internal enforcement of terms and conditions, processes and procedures, use of automated tools, risk assessments, and user education efforts (Box 17). These transparency reports will be publicly accessible (4.20).

<table>
<tr><th>Out of Scope</th><th colspan="2">Within Scope</th></tr>
<tr><td rowspan="3">• ISPs
• B2B services
• Hosting providers
• App stores</td><td colspan="2">Scope includes search engines, as well as companies who:
• ‘host user-generated content which can be accessed by users in the UK’ (and/or)
• ‘facilitate public or private online interaction between service users’

Obligations:
• Overarching duty of care to ‘prevent user-generated content or activity on their services causing significant physical or psychological harm to individuals’
• Take action on relevant illegal content
• Assess if children likely to use service, and provide protections if so.

Exception:
• Robust protections for journalistic content shared on in-scope services</td></tr>
<tr><td>Category 1</td><td>Category 2</td></tr>
<tr><td>Additional obligations to comply with priority harms including:
1. priority criminal offences e.g. CSEA, terrorism, hate crime;
2. those which are harmful to children e.g. pornography;
3. that which is harmful to adults e.g. abuse, eating disorders, self-harm.
– Transparency obligations</td><td>No additional obligations</td></tr>
</table>

Ofcom will have **information gathering powers**, including the power to interviewee employees and to ‘enter companies’ premises and access documentation, data and equipment’ (4.26). Additionally, Ofcom will have the power to require a company to undertake an external ‘skilled person report’ which would have particular usefulness where external technical expertise is needed, for example in validating ‘the effectiveness of automated moderation systems’. (4.28)

Ofcom will be required to produce a report on **independent researcher access to data**, assessing the ‘opportunities, challenges and practicalities of companies providing independent researchers with access to company data’ (4.29).

Enforcement powers include issuing directions for improvement, notices of non-compliance (4.43), as well as sanctions in the form of civil fines up to £18 million or 10% of annual global turnover, whichever is higher (4.43). In cases of egregious non-compliance Ofcom will be able to take measures to disrupt business activities, including blocking access (Box 19). These enforcement powers are intended to be used against companies with and without a physical or legal presence in the UK.

Oversight & Accountability

Ofcom will present annual reports before Parliament and is subject to Select Committee scrutiny (3.17). Parliament will approve all codes of practice that Ofcom produces (3.18). A review of the effectiveness of the regime will take place 2-5 years after its entry into force (3.19). Ofcom will conduct and publish impact assessments for proposals which will affect business (3.20).

Endnotes

1. https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN
2. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348
3. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348
4. DSA, p.4.
5. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348
6. For a complete list of relevant sector rules see Article 1.5.
7. This principle does not affect the possibility of injunctions against providers by courts or administrative authorities in the member states (§24).
8. Number in brackets refers to the relevant paragraph in the full government response to the consultation. See fn. 9
9. Online Harms White Paper: Full government response to the consultation. https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response#part-1-who-will-the-new-regulatory-framework-apply-to
10. https://www.gov.uk/government/publications/online-harms-interim-codes-of-practice

ISD | Powering solutions to extremism and polarisation

Digital Policy Lab
Provocation Paper

The Liberal Democratic Internet – Five Models for a Digital Future

Alex Krasodomski-Jones

About This Paper

This provocation paper explores proposed settlements on the balance of power and what they mean for the future of the web. It highlights the ways state, corporate, individual and machine power might help or hinder the democratic project, and the balance of powers proposed by competing conceptions of government.

The views expressed in this paper are those of the author and do not necessarily reflect the views of Digital Policy Lab participants or governments.

About The Author

Alex Krasodomski-Jones is director of the Centre for the Analysis of Social Media, co-lead by Demos and the University of Sussex. Demos is an independent, educational charity, registered in England and Wales (Charity Registration no. 1042046).

Alex has authored more than a dozen major reports on digital election integrity, content moderation practices, digital regulation and the intersection between tech and politics. He leads the centre's visual analytics practice, and provides written and televised comment for the BBC, CNN, Wired, the Spectator and other outlets.

Alex leads Demos's effort in the Good Web Project, a coalition of civil society organisations working to articulate, measure and advocate for an Internet governed on open, liberal and democratic principles. His focus is on building an evidence base, both for what good practice looks like in the architecture and practices of future online spaces, and of public and government consensus around that future.

ISD | Powering solutions to extremism and polarisation

DEMOS

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Summary

Without a principled vision for the internet, our democratic traditions, values, government and society risk falling behind authoritarian states, technopolistic industry giants and autonomous technology in the race to reshape the most important international political, cultural and social space in existence.

The web was born from circumstance: a US project, then a Western project, then a global one. An academic experiment with military origins built first by digital visionaries, before power passed to the Internet giants with whom our lives are now totally intertwined. This origin story is retold in the principles that underpin the web as we know it.

Those principles are now in question. At first, it was authoritarian regimes that were wary. Now, the world over, governments are vying for change. The future of the Internet is in doubt, and no cohesive settlement has been found.

The balances of power between states and corporations, corporations and citizens, and the social contract between states and their citizens is in constant flux online. Powerful technologies – artificial intelligence and trustless technology – present a fourth pressure, with our lives increasingly governed by machine, not man.

This short paper explores proposed settlements on the balance of power and what they mean for the future of the web. It highlights the ways state, corporate, individual and machine power might help or hinder the democratic project, and the balance of powers proposed by competing conceptions of government. The paper demands we reset our vision for liberal democracy in a digital age at this juncture, to win over our publics to a vision of something better, and to secure that vision in collaboration with our friends and partners.

Where we are, and how we got here

No technology could claim greater political importance than that which underpins the Internet. Technology is not created or deployed in a vacuum: it is inherently political, with social and cultural contexts shaping what technology is built, how it is designed, and to what ends it is used.

To understand how we have got where we are, we can peel back the layers of information and take a closer look at the technology that ferries it all around. These are the web's protocols, an intimidating-sounding name used to describe the rules for how information is communicated.

We have offline protocols of communication. We might shake hands, for instance, or kiss each other on the cheek (once in Peru, twice in Croatia, three times in Belgium) to say hello. Examining these habits can tell us something about a society, just as looking at those protocols online can tell us something about the principles of the Internet as it stands.

The Internet Protocol Suite (IPS), often referred to as TCP/IP (Transmission Control Protocol/Internet Protocol) after its two most well-known protocols, bears the political and cultural scars of its genesis. The web's origin story is one of squabbles and conflict over its design: over the level of central control possible, over the scale of military and government involvement, and over who should reap its rewards.[1, 2] Paul Baran's early projects building resilient global communication at RAND were set against a backdrop of the Cold War, and the eventual ARPANET that emerged in the 1960s was built by academics and funded with military money. States, companies and institutions vied for power while the network's growth accelerated. Corporations then struck gold. Soon after the dot com bubble burst, Google, then Facebook and others discovered the optimal business model for making money through this technology. It took fewer than twenty years for their applications to form a new bedrock for global business, culture, society and politics.

The larger it got the more deeply embedded the protocols, principles and norms on which it was built were buried. They reshaped the world.

This network's protocols support a "dumb, trusting middle" with "smart, anonymous hosts on the edges".[3] IP itself neither cares nor can control what information is sent across it, nor who sends or receives that information: a barrier against centralised control or surveillance. It is trusting. It expects good behaviour, with limited recourse for when people break customs or rules, or abuse it. End users are empowered, anonymous to most participants, and free to join as they wish: scalability is prioritised over any centralised control. In her history of TCP/IP, Rebekah Larsen tells the story of Vint Cerf's switch: one of the fathers of the web had an on/off button for the entire network used to force through updates to the original ARPANET. We are a long way away from that kind of control now.

Anonymous, free, open, trusting, decentralised and resistant to central control: these are the founding principles of the web as written in the technology that knits it together.

How comfortably these principles sit with states and citizens is in constant flux. For some people, at some moments, some are welcome. A Silicon Valley entrepreneur during the dot com bubble or a sympathetic onlooker during the 2011 Arab Spring might have celebrated what they saw the Internet as enabling, but so might a scammer or terrorist recruiter. A dictator wary of their citizens' freedom of speech, freedom of press or freedom of association may have been more worried. So might parents concerned about their children's browsing habits or security officials facing new forms of information warfare and digitally-enabled crime.

These concerns define the struggle for the future of the web. It is being fought in shareholder meetings and across front pages, in impenetrable tech roundtables and in the homes of each and every Internet user, in every arena of government activity: investment, war, trade, regulation, security-provision and so on.

One compelling narrative is found in Wendy Hall and Kieron O'Hara's seminal Four Internets paper, which tells the stories of these perspectives.[4] It describes Silicon Valley's Open Internet, the grandchild of the early web where technology and profit drive innovation and principles of freedom and unfettered access remain, albeit caveated by commercial imperatives. Under this model, the state comes second to corporations and technology in determining the rules of the game.

By contrast, "Beijing's Authoritarian Internet" is ideologically positioned as a tool of surveillance and control. States like China are fed up of the "dumb, trusting middle" that acts as a bulwark against government surveillance - benign or otherwise. Private corporations are extensions of the state. Chinese web giants answer to the government, not the other way around.

Alongside these monoliths lie visions that run the gamut of political ideologies: liberty absolutists, for instance, who demand the removal of even existing regulation or protocols deemed anti-competitive. This vision for the web is embodied in Four Internets by Republicans in DC, but perhaps is also found in the more visionary crypto-anarchist and alt-tech world hell-bent on maximising individual liberty over responsibility.

Against this backdrop of warring visions, a new power is rising: self-determining technology – tech that automates decisions, or even writes its own rules. We are not at the singularity quite yet.[5] Nevertheless, governments and corporations are wholeheartedly backing machines in making decisions for them, from cancer screenings to exam results, from the news we should read to the stuff we should buy, decisions outside of standards of access, transparency or redress. The maths behind end-to-end encryption minimises the trust a user needs to put in central authorities much like the principles underpinning public blockchain technology. A user puts faith in the technology, not in government or society or a corporation. Wherever we see these entities and institutions struggling with technology, it may be time to stop and question whether it is the technology itself that is challenging their power.

The principles of the web as it has been developed have been confusing to liberal democracies. Some are welcome: an expectation of civility – some might say naïve – can be found in both liberal democracies and in the architects of the web's foundations. Rights to privacy, freedoms of speech and of association are founding principles of liberal democracies, and we celebrated the sight of smartphones in Tahrir Square.

Other principles have become cause for concern. Democracies turn on a social contract, a trust in government, and the web has repeatedly tested the limits of states' power to have their way. The "dumb middle" and subsequent privacy-boosting technologies (such as the uptake of encryption) challenge the state's ability to carry out one of its fundamental duties, namely the preservation of its citizens' security. This comes alongside simultaneous concerns about the exploitation of citizens' privacy by companies often outside their geographic jurisdictions.[6]

Taken together, we identify four forces able to shape the Internet going forward, four powers to whom we must assign responsibility for digital life. These powers are states, corporations, individuals and machines.

States, Corporations, Individuals and Machines

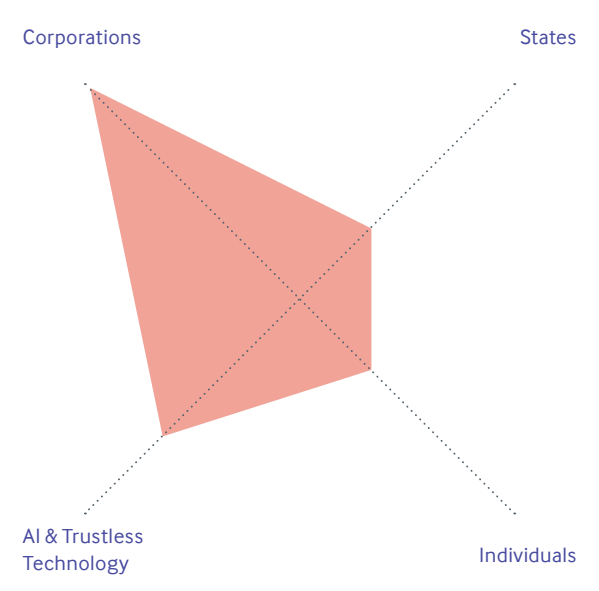
Four powers will be responsible for the shape and quality of the Internet: states, corporations, individuals and machines: AI and trustless technology.

State control of the Internet can take many forms, but in this picture includes efforts to regulate and control the shape of the web by national governments and international cooperation. Under a democracy, it is the rule of law and its enforcement.[7] What that looks like in China is very different to what it might look like in Europe, or under a UN or other international treaty, but in all aspects it involves subjecting the Internet and its underlying technology to rules drawn up by governments.

Corporate control is different. Here, companies write the rules. Whether it is the speed or breadth of content moderation policies or the nature of the content that filters through to each user, the rules and mechanisms governing those processes are defined in boardrooms before legislatures. The relationship between state and corporate control is complicated. In some cases, states may devolve rulemaking to companies on the grounds that these are private entities who have the right to set their own standards. In others, governments may simply lack the power or jurisdiction to compel or coerce a platform into changing, either because a platform is unwilling or because the platform simply cannot carry out whatever it is being asked of it. The ongoing fight over copyright content is a useful example of this: some platforms are unwilling to remove copyrighted content, while others lack the technology to detect and remove it quickly enough. In both cases, the state's power is secondary.

Individual power foregrounds citizens' responsibility and ability to meaningfully understand, influence and control the online world around them. The informal agreements written into the protocols of the early Internet are clues that its architects put a high premium on the freedom and power of its users. Giving power to users to better manage their personal data and cultivate the spaces they use online is core to this. Ensuring people have autonomy online has not been a priority of the corporatized Internet as we know it, with its economic model of targeted advertising so dependent on data extraction and a compliant user base. Protection from harm and equality of opportunity are at best lines on a balance sheet.

Silicon Valley



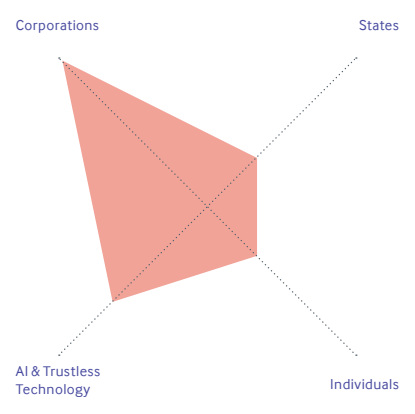
A sketch of power balances under one model, the technocentric, corporate Silicon Valley model.

Finally, there is **machine control**, or more accurately artificial intelligence and trustless technology like encryption, blockchain and cryptocurrencies. Although these technologies are all different, they share a commonality: it is not a human that sets or enforces the rules, but a technology. An AI can diagnose cancer, a bitcoin can be bought or sold or traded, and a deed transferred on a blockchain all without the oversight of a central authority. In the past few years we have already seen the growing power of these technologies to shape our lives, and their existence implies decisions outside of any corporate or state view. The maths behind end-to-end encryption is public knowledge, theoretically open for anyone to use and near-impossible to censor or ban. But its use creates channels that are by definition not accessible to states, corporations, or other individuals. Blockchain technology is frequently touted as an exercise in removing state and corporate control from a system: Bitcoin needs no central bank. Artificial intelligences, even those nominally in the hands of states or corporations, are frequently so complicated that their decisions cannot be simply explained, computed or reverse-engineered. Citizens already cede decision-making control to algorithms every day, when shopping or navigating, and governments around the world are increasingly turning to algorithms to make decisions. States and corporations may believe that AI is little more than an extension of their own rule-making power, but this is short-sighted. First come the AIs that civil servants and marketing executives do not understand but deploy anyway. Second come AIs so complicated that even their creators cannot fathom quite how they work. Thirdly, and finally, come future AIs powerful enough to write their own rules and carry out government activities far more effectively than any human organisation and are consequently resistant to oversight, accountability or explanation.

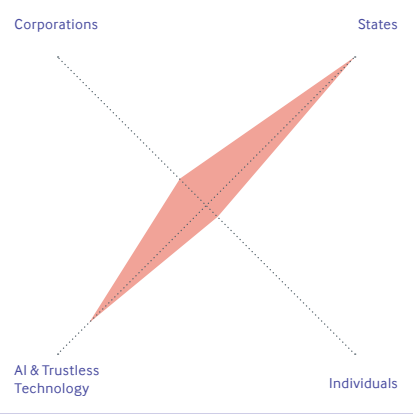
Machine power is not a spectre: even in 2021, billions of people are subjected to decisions made by machine in all aspects of their lives. Ignoring this power and failing to regulate its use would be a mistake, and early efforts to this end include ongoing work on ethical AI and ethical use of AI, as well as in protests against its use in courts, policing and education.

The lines are blurred. Corporate control may dissuade government oversight by implementing encryption, as Facebook is threatening to do, while states may incubate compliant corporate players like Sina Weibo or WeChat in China. Nevertheless, these four types of power offer us useful pivots around which to imagine a future Internet.

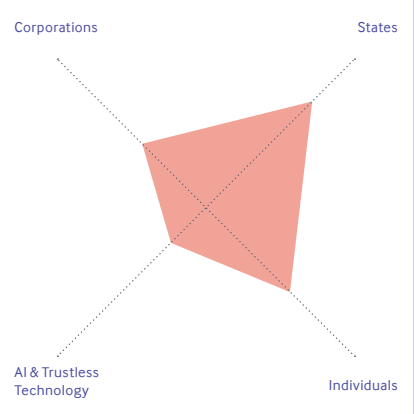
Where shoult Power Lie?
Six models for a Digital Future



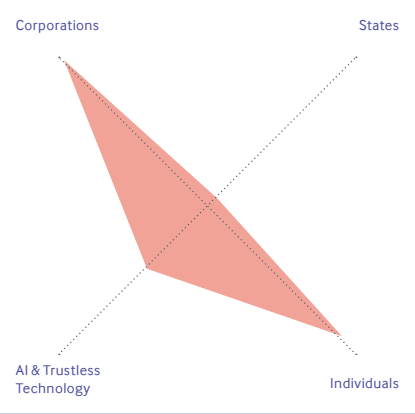
A Silicon Valley Internet



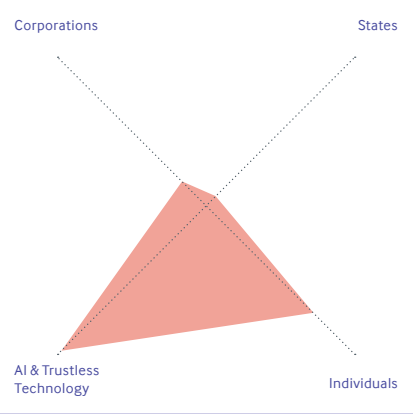
An Autocratic Internet



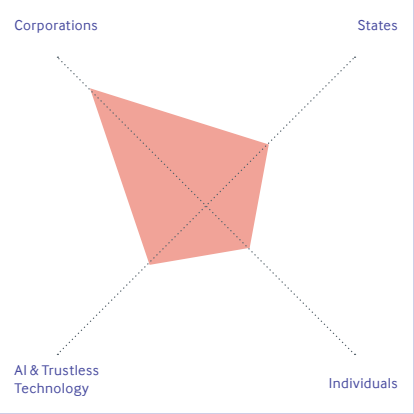
An EU Internet



A Libertarian Internet



A Machine Internet



The UK in 2021

The Shape of Things to Come

These diagrams are caricatures, but show how different the competing visions for the Internet might look. Each presents its own set of threats and opportunities.

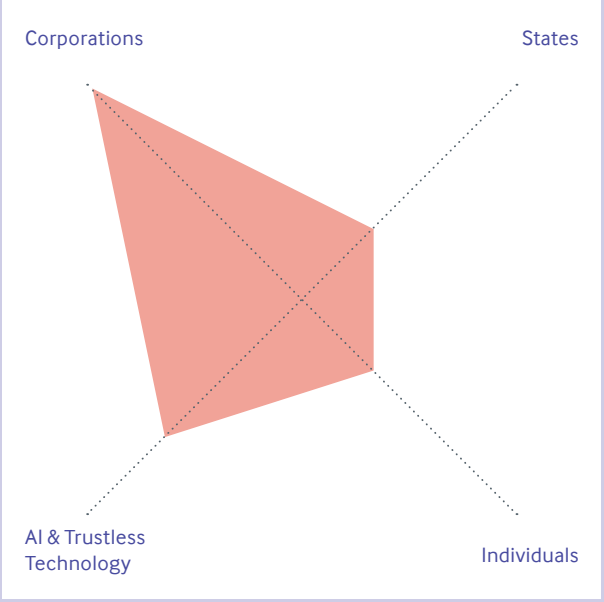
The Corporate Internet

The corporate Internet is the closest to the Western world's status quo. Under this model, it is the major international corporations that wield the greatest influence in determining the shape, cultures and rules online: whether Facebook, Amazon or Google and its subsidiary YouTube. Outside of a narrow band of illegal content, the limits of free expression are determined in Silicon Valley boardrooms or across the constellation of smaller platforms sustained by advertising revenue, the mechanics of which are frequently provided by the same technology companies.

Tension exists between the infrastructure providers and the applications that run on them, but under this model tension is resolved through the private sector. The expansion of Tesla's Starlink programme as a corporate-owned, international provider of Internet access is a useful indicator of what is to come, with corporations bypassing state-imposed infrastructural limits on their activity.

The relationship between platform and state is a one-sided negotiation. Regulation moves slowly, is continually contested, and application of the law is frustrated by a lack of transparency and meaningful ways to measure or survey what is happening on these platforms at any one time. Government data access and collection is feeble when compared to the powers of the platforms. Individual users fare even more poorly: the services offered are extraordinary and nominally free, but are exchanged on terms that utterly disempower their users. Redress, control or engagement on platforms is little more than a veneer concealing this asymmetry-by-design.

Silicon Valley



Technology plays a key role here. Encryption frustrates oversight, and is deployed as much to protect market share through security as it is to create distance between the platform and the content circulating on it. AI and algorithmic curation is the only feasible route to managing spaces this large and to maximise data capture and advertising revenue, and the functionality of these algorithms is opaque, their decisions broadly unchallengeable.

The Autocratic Internet

The shape of the online world evolving in China (and under its growing sphere of influence in the developing world) stands in contrast to the 'technopoly' of the corporate Internet we are familiar with in the West. Here, the state calls all the shots, and platform technology is an extension of the government's power rather than a thorn in its side. Power invested in individuals is minimal.

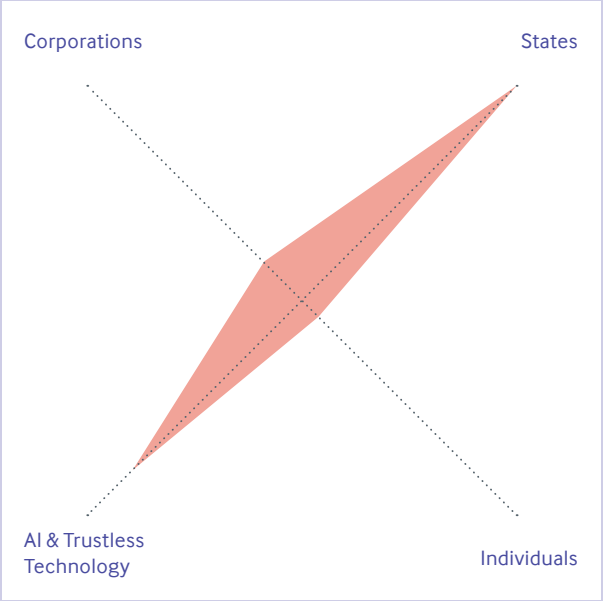
Protocols and infrastructure are state-centric and prioritise state sovereignty. At their most sophisticated, they include hard limits to the boundaries of the web, like the Internet found in North Korea. At their crudest, they are an on/off switch.

Individuals' rights remain limited. Under this model data access by government is facilitated by platforms, and enormous, joined-up data on Internet users underpin surveillance, social score systems and experimental technology. The subjugation of Uighurs in China is facilitated by technology above all else.[8]

This same data unlocks the full potential of state-aligned artificial intelligences, themselves a further extension of state power in as far as their outputs and operations remain knowable and intelligible. Citizens' rights of redress are negligible, regardless of whether a decision is made by an AI or the state, and that difference will become increasingly blurry.

Harnessing machines in service of the surveillance panopticon means stamping out some technology as much as encouraging others. China's 2020 encryption law introduces a tiered approach which critics describe as tantamount to the ban on end-to-end encryption for everyone but the ruling party.[9] Cryptographic applications like Tor, Telegram, WhatsApp, Mastodon or Virtual Private Networks (VPNs) are banned in the country.

An Autocratic Internet



A Libertarian Internet

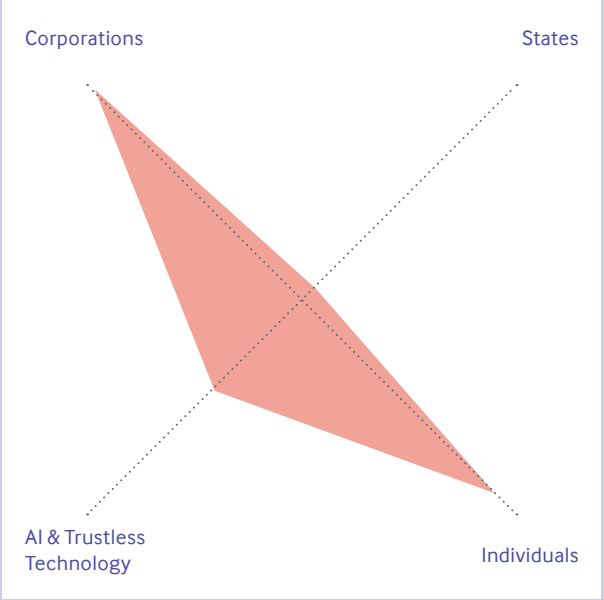
The US position on the future of the Internet is often associated with ideas and ambitions of the web giants that call it home. But there is division in the country, and a competing vision for the soul of the web: DC's commercial Internet, an Internet prioritising private actors from platform to infrastructure provision, a market free from any regulation whatsoever. Freedom of expression in this model is interpreted as freedom from state intervention, rather than state-guaranteed equality of opportunity.

Under this model, there should be nothing stopping an individual from creating, accessing or participating in digital services online, and individuals take responsibility for their behaviour under terms set and enforced by other individuals. Protecting one's privacy or rights online falls to the individual. Limits on freedom, such as legal codes of speech or expression, are anathema here, as are rules demanding equality of opportunity.

Under this property-based model, corporations have no expectation of providing anything save from what their customers might want. Internet Service Providers (ISPs) should have the power to maximise their profits, and long-standing web principles like net neutrality stand in the way of this. Under this model, there is no expectation of public good or openness in platforms, nor is interoperability between sites and services necessarily supported. This runs contrary to the hopes of early Internet pioneers, for whom the web should be a single, connected information space. Instead, the Internet is fragmented into profit-driven "walled gardens".[10]

Power here is shared by corporations and their customers, free from state oversight. It is a model invoked by those who reject government intervention at a more microscopic level. Pressure on major platforms to reduce online harms has led to the proliferation of so-called 'free speech havens', alternative technology platforms like Parler and Gab that tend to cater to extremist political positions nominally banned by the Silicon Valley giants' terms of use. However, the infrastructural weaknesses of these alternatives has been brought to light in early 2021, with Amazon Web Services' suspension of Parler following the attacks on

A Libertarian Internet



the US Capitol. Under the Libertarian Internet model, these spaces are promoted: there is a market for them, and so they ought to be allowed to satisfy that market. Pressure on service providers to censor these spaces will accelerate their growth and distribution.

Trustless technology under this model becomes just another product feature. If customers demand security, there should be no barriers to implementing powerful encryption to your service if that is the route to maximising your customer base and outcompeting competitor services.

A Machine Internet

Finally, we outline a fourth framework: A machine Internet where politics, society and culture is governed by rules set by artificial intelligences and code.

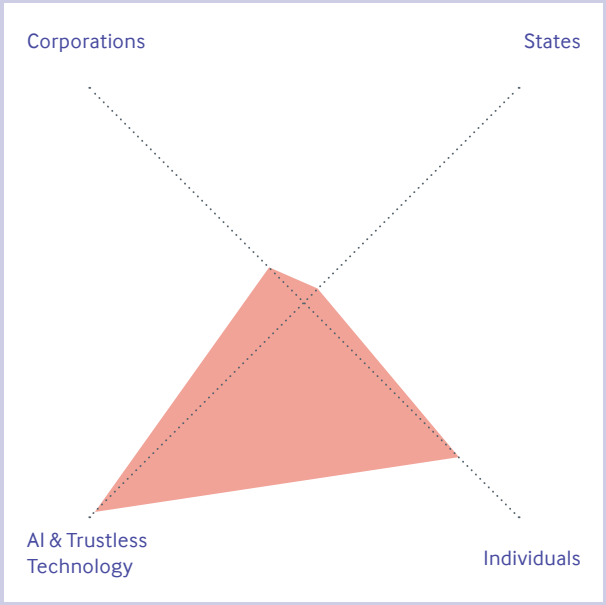
In this conception, the technology itself sets and enforces the rules: at first at the behest of a state or corporation, but eventually outside of any corporate or state interest. This is the vision of the Internet furthest from our current one, but the growth in machine-enabled decision-making and crypto assets make a world where code is law worth exploring.

Governance by AI is on the rise. Sufficiently powerful AIs will be employed to make decisions about increasing parts of our lives, beginning with the routes we take to work, through our ability to access credit or buy a house, and eventually culminating in AI-enabled law enforcement, national security, and provision of public services.

Challenging decisions made by algorithms is already difficult given the level of technical expertise required. Computational decision-making has been shown to be more effective than human decision-making in some domains - in diagnosing health conditions, for instance, or in identifying fraud. Under a machine internet, AI systems are expected to replicate the functions of government more effectively and efficiently, eventually replacing them piece by piece. This has clear and unresolved ramifications for questions of democratic choice and political representation. Biased, opaque algorithms fail any democratic test.

Again, this is not science-fiction. Every day, billions of people globally are subjected to decisions made by machines that they do not understand nor have any power over or expectation of redress. Every day, governments come face-to-face with technology that limits their power.

Uptake in cryptocurrencies like Bitcoin has tended to be driven by speculation, but the use of cryptocurrencies is seen as a route towards providing financial services to people who fundamentally distrust corporate or central authorities. New, so-called permissionless systems, digital autonomous organisations (DAOs), smart contracts and so on are all built to allow the transfer



of money and commercial cooperation among users entirely without third-party involvement or oversight, be that corporate or state.[11] The architects behind these systems imagine a world where digital technology replaces nation states by enabling individuals to cooperate through technology alone.

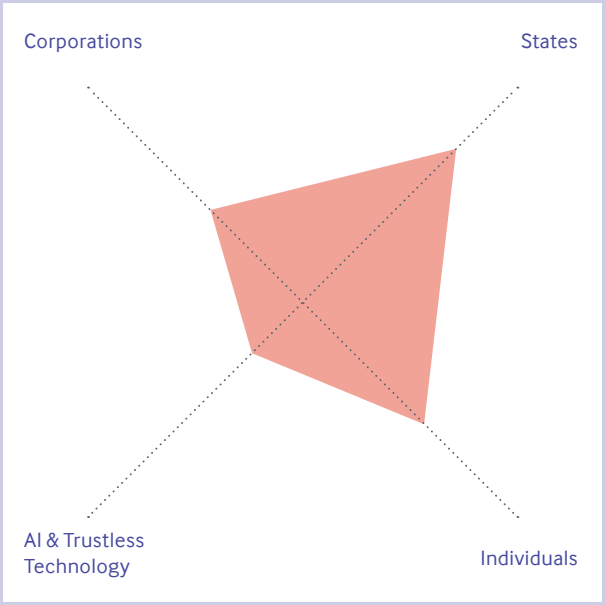
The EU Internet: A Liberal Democratic Internet?

This model - Hall and O'Hara's Bourgeois Internet - is best described as the state fighting back. It may also be the closest governments around the world have got to a liberal democratic web. Digital regulation led by the EU and its member states is reactive, and driven by attempts to remedy perceived harms and threats enabled by the corporate Internet. Although increasingly couched in proactive language, EU regulation has primarily been remedial. General Data Protection Regulation (GDPR), the Google Spain vs AEPD case, the NetzDG laws in Germany and most recently the EU's Digital Services Act (DSA) are useful examples of states taking steps to curb the behaviour or design of (primarily US) technology platforms.

A vision is now emerging for what this Bourgeois Internet might look like. In this imagination, state power is deployed as the key defence of citizens' rights and liberties, and citizens are expected to put faith and trust in national and international institutions. It places a heavy emphasis on the role of citizens, trusting them, in return, to act with civility and tolerance. Data rights are better protected, with a trajectory towards greater citizen control over the use and value of the data they produce.

Untrammelled machine power presents a threat to state hegemony, and this is as apparent in the Bourgeois Internet as anywhere else. The EU has led the pack in calling for ethical standards for artificial intelligence, recognising the increasing use of this form of decision-making. Civil society organisations are vocal in calling for algorithmic transparency, rights of redress, and for caution in the implementation of AI-enhanced technologies like facial recognition. While the roll-out of trustless technology has been tolerated, laws around the advertising and provision of cryptocurrency services have been implemented. The debate over privacy-enhancing technologies like end-to-end encryption continues to demand a settlement: a dilemma between safety and security on the one hand and rights to privacy, freedom of expression and commercial questions on the other.

An EU Internet



The Liberal Democratic Internet

Assigning power to states, corporations, individuals and machines all present both threats and opportunities to liberal democratic development. Navigating these ambiguities and dilemmas won't be easy, and time is short. The moment for celebrating the web as a powerful tool in projecting liberal values is over: it was never inevitable, never the End of History. Managing speech and information in a liberal democratic society is a painstaking exercise in slow-moving regulation, care and caution. Timidity and patience is easily exploited in the fast-moving world of technology.

The mission for liberal democracies, and that of the Good Web Project, will be to **identify the technologies, design principles and governance that ensure a balance of powers commensurate with liberal democratic values**. The breadth and depth of the challenge is formidable.

There is work to do across every layer of the technology stack that makes up the web, from individuals' rights and liberties up to scales as grand as international security and sovereignty.

In the sections below, we map these dilemmas, identifying the threats and opportunities across three broad areas: **the digital citizen, the digital commons, and security and sovereignty**. For each, we identify where liberal democracies ought to step up their defence and support, and where the threats from corporate, state, individual or machine power require particular vigilance.

Digital Citizenship

The defence and promotion of citizens' rights and liberties online, and their active participation in online life, is the foundational challenge facing liberal democracies as they look to reshape the online world.

It is barely an exaggeration to describe the democratic disempowerment of the average user online as the Internet's greatest tragedy. In most Western countries, the active participation of citizens in political and civic life has been utterly subsumed by the prerogatives of monopoly platforms and the economic model that underpins their design. The average Internet user has no power to reshape or cultivate the spaces they live in, limited as they are by arbitrary, confusing or inconsistent terms of use and platform enforcement. The ability to choose those that govern us is a core tenet of liberal democracy, but online users have no right nor route to contest the decisions made by higher powers under the default platform model. "Within this framework," writes Giovanni De Gregorio, "the lack of any users' rights or remedy leads online platforms to exercise the same discretion of an absolute power over its community."[12] Shoshana Zuboff calls these "the social relations of a pre-modern absolutist authority".[13] Others have called the platform model feudalistic or Hobbesian: a system under which you give up your rights in exchange for products and services.[14] Whatever it is, the current situation does not sit comfortably with our conception of citizens in a democracy.

Defence of citizen rights and responsibilities by states, along the lines of a traditional social contract, has been frustrated by corporate power in liberal democracies, and was never a prospect under authoritarian regimes. The COVID-19 pandemic has shone a light on the dangers presented by the digital world when state power is unfettered: AI-driven cameras, data capture and analytics, and facial recognition software ensure citizens are carefully monitored, and infractions against a law or directive are significantly more likely to be detected.[15]

The rule of law itself has been weakened: enforcement is harmed by patchy capability, out-of-date legislation, limited access to evidence and weak international coordination. Moreover, the online boundaries of acceptable behaviour are shaped by terms and conditions long before law and its enforcement.

Finally, the rise of machine-enabled decision-making presents new threats to traditional conceptions of citizen power. Already, trustless technology like end-to-end encryption has by design rewritten the rules on human rights: at once a boon and a risk to rights to privacy and security, while also extending new powers to a select group of technologically-savvy individuals.

Code becomes law. Our lawmakers are first engineers, then artificial intelligences. The routes to political and social participation and the rights and freedoms of participants will be defined not through human oversight, but by the technology itself, ushering in new questions for how humans can wield power in a world of machines.

Bringing these forces instead to the defence of the rights of citizens and to the service of citizen empowerment is paramount. Ensuring corporate power is checked by law and government power is an essential first step to ensure citizens' rights are defended, and that citizens are able to comprehend, affect and challenge the spaces they live in online. While web giants may be the target now, the same questions must be applied to machine power in the future, ensuring algorithms and artificial intelligences operate along lines consistent with liberal democratic principles. Strengthening the power of communities online is an essential step, and the Internet is made up of thousands of examples of how to do this effectively.

The reward of a properly balanced system will be the practical application of these powers in the service of citizen empowerment. State power, and the rule of law, should protect citizens against corporate misdemeanour, and ensure that rights, responsibilities and liberties are enabled by digital design. Corporations, properly empowered, will develop competing models for online life, providing citizens with genuine choices over where they live their online lives, and bring liberal democratic technologies to new audiences around the world. Citizens properly empowered to take responsibility for their online lives will find routes towards a meaningful digital civic society, forming and

cultivating new communities and relationships on the terms of their choosing. Trustless technology can be deployed to protect rights and liberties in environments where autocratic states and corporations abuse their power. Blockchain technology has already been used to protect rights to property in places where that right is less than guaranteed.[16, 17] Carefully designed artificial intelligences may well increase citizen capability to live full and free lives through improvements to decision-making, information access and new models of work and social support. The ethical use of AI has frequently been touted as an area where liberal democracies may have an edge.[18]

Case Study

Facial Recognition

Without sufficient power for citizens, technology can be all too easily weaponised by states and corporations against individuals and communities. The use of digital surveillance by the Chinese government to perpetrate atrocities against the Uighur people in Xinjiang has long been recorded.[19] Recently it was revealed that the company Huawei had been involved in testing AI facial recognition technologies to identify people's ethnicities which could send a 'Uighur alarm' to the police if a member of the minority group was identified.[20]

Yet the wielding of power by corporations abetting state oppression is a global concern. Some corporations have tried to distance themselves: after it was revealed that US law enforcements' use of facial recognition systems displayed severe gender and racial biases, Amazon, Microsoft and IBM ceased sales of such technology to those entities..[21, 22]

The use of various facial recognition systems by law enforcement and private companies has also been the subject of lawsuits from South Wales[23] to Illinois.[24]

We are in a situation where state power - and state repression - can be amplified on a huge scale through the use of unaccountable technologies: where we rely on the goodwill or reputation of companies, or, where they exist, cumbersome legal processes to constrain abuses. A liberal democratic settlement cannot be content with always playing catch-up to the relentless pace of technology developed and adopted, with democratic oversight only ever, if ever, an afterthought.

A Digital Commons

We entered the 21st century with a series of assumptions about what makes for a 'democratic' information and media space and supported a full, free and fair public debate: freedom of expression; pluralism; the metaphor of a marketplace of ideas. But online spaces have frequently failed to meet these ideals, and pose a new challenge to the coherence of those original principles.

The common analogy is that of the shopping centre or mall: areas that feel like public spaces in the offline world, but have their own rules, drawn up in private, and enforced by private security. The centralisation and homogenisation of digital public space by a handful of US companies has left the design, the cultural norms and the shape of the public discourse enabled by media in the hands of corporate power. The design imperatives behind these spaces are clear: maximising shareholder values requires a panopticon of data collection and the prioritisation of attention-grabbing content. These spaces are at once rigidly controlled in defence of those powerful or wealthy enough to maximise their share of voice, and simultaneously exploitable by actors savvy enough to do so. Public service media increasingly resembles a model dependent on charity.

Machines play an integral dual role in maintaining this control. They serve to maintain the enormous private commons represented by social media platforms. Complicated algorithms prioritise content for profit, clunky algorithms censor speech and information automatically, personalisation algorithms segment and tailor information to the point where two citizens might live in utterly divergent realities.

Democratic states remain firmly on the back foot. As major funders of platform advertising, states have found a route to make the most of this new world order without meaningfully challenging its principles, and attempts to preserve the principles of public space have been limited to reactive regulations targeting online harms. Solutions are not easy, and well- meaning attempts at digital regulation have often

Case Study

The Section 230 Conundrum

One of the apparent oddities about the 2020 US Election was how former President Trump and President Biden could come from such different positions on what the online world should look like, and both arrive at the same conclusion: that Section 230, which protects internet companies from liability for content hosted on their platforms, needed to be reformed.[25] Trump's longstanding (unevidenced) complaint against the big tech companies has been that, in taking action against hate speech and extremism on their platforms, they 'censor' conservative and right-wing voices.[26] Biden has said, conversely, that the proliferation of misinformation and disinformation online is cause for reconsidering the protections.[27]

What is clear is that without a common vision of what a 'good' public space actually looks like (no misinformation? no censorship?) approaches to addressing the power imbalance online will be piecemeal and inconsistent. Moreover, the ownership of these spaces by private companies who operate without oversight or significant transparency means that these kinds of contradictory conclusions are likely, as government fights to get back power however it can: whether or not it actually succeeds in enfranchising citizens.

succumbed to an authoritarian vocabulary of take-downs, blocks, bans and censorship.

Authoritarian states, by contrast, have made the most of the digitisation of public space, either by piggy-backing on the surveillance machine or by exploiting its weaknesses within their borders and without.

Individual power in shaping public spaces is incredibly limited when the world is one great shopping centre. Creating or maintaining public space on the Internet is a thankless task for those not able to monetise it. The handful of people able to sustain an online presence as a commentator, journalist, public figure or talking head do so sharecropping through Premium Snapchats, on Amazon's Twitch or Google's YouTube.

Redrawing the public sphere must be a critical priority for democracies, and a liberal democratic Internet requires change at all four corners of the sketches above.

Corporate provision of public space, usually expanded under the proviso of connecting the world, could act as a powerful arena for the projection of democratic values. But new models for sustaining public space and the voices within it are vital. Regulation in favour of alternative models of public media and the restoration and preservation of funding models outside of advertising revenue are vital routes towards ensuring media is plural, responsible and sustainable.

Unaccountable and opaque machine power cannot be entrusted with the governance of the digital commons. Where space is necessarily maintained by algorithms, changing their design and boosting their transparency should build a technologically-enabled public sphere where machines are deployed in defence of minority voices and the preservation of a free and open media. Properly deployed and managed, technology like encryption and decentralised networks can serve as a further line of defence for the public sphere in the face of those who would look to see it surveilled, controlled or shut down.

Empowered citizens are custodians and participants in the digital commons. With the right incentives, a liberal democratic Internet will see the transformation of its users from digital serfs to digital citizens, empowered to shape and contribute to a healthy and vibrant public sphere.

Security and Geopolitics

Over the past five years, the cold war online turned hot. Battles over digital sovereignty began with domestic developments, with nation states like Russia and China pressing for greater control over the web within their borders. The Internet has become a vector for international geopolitical aims, too: both through the weaponisation of open, online spaces and the deployment of disinformation, and through the race to deploy digital infrastructure around the world. Running concurrently with these grander plans, cybercrime is the fastest growing threat to citizens: from scams and identity theft to extremist recruitment and the marketing of child sexual exploitation online. Liberal democracies have been slow to respond to these threats.

Corporate power, embodied in the policy and resilience teams inside major platforms, has been exposed. Platforms were asleep at the wheel: either unaware of the ways their services were being exploited, unable to counter it, or choosing to ignore it.

Individuals have been reduced to cannon fodder. Forced into the front line by platforms hell-bent on connectivity and growth and lacking digital literacy, they have been easy prey for groups and individuals looking to exploit them. Education initiatives and fact-checkers were orders of magnitude too weak to be viable tools of self-defence. Fraud and cybercrime is thought to affect one in three Americans.[28]

The state's ability to protect its citizens has been called into question time and again as our lives move online, with encrypted devices and communication platforms adding a further barrier to law enforcement tasked with tackling digitally-enabled harm. As noted above, technology that is resistant to centralised control and oversight inevitably limits the power of central state or corporate authority.

Nation states relying on reactive regulation as a tool to combat the influence of platforms have also been slow, powerless to defend the new information landscape and its haphazard Silicon Valley custodians against foreign actors. A lackadaisical approach to infrastructural development has seen countries reliant on imports from authoritarian regimes in their own backyards. There is a vacuum in competitive infrastructural offerings in the international market when compared with the scale and ambition of China's Belt and Road Initiative, for instance. As we enter the age of the Internet of Things, there continue to be questions about the security implications of the devices being sold to millions.

International cyber supremacy will be dictated in major part by machine power. In the hands of states and corporations, this means the development of artificial intelligence. As noted in Four Internets, the ability by authoritarian regimes to bypass concerns over data privacy and amass enormous, connected datasets on which to train AI may give them an advantage in developing superior products. Chinese-owned apps like TikTok are already finding Western audiences while Silicon Valley applications are banned or neutered within Chinese borders.

Democracies need to define a liberal doctrine of security and sovereignty, one that recognises the threats caused by information operations and cyberattacks, foreign and domestic, as well as online crime, but also guarantees freedoms and the free flow of information across borders.

Empowering states and multilateral institutions to secure and defend an open Internet is a vital step in reasserting sovereignty in the online world. This requires change and improvement to the network architecture of the web, both to reinforce the open Internet in the face of protocols designed to fragment it, and to ensure that liberal democratic principles continue to be reflected in the underlying technology. Improved transparency of digital standards bodies and involvement by multilateral institutions must be mainstreamed. Where corporate monopolies are identified as a weakness in national and international security, those weaknesses must be addressed, ensuring global corporations are a vanguard of liberal democratic values instead of undermining them.

Further, states must move beyond authoritarian vocabulary of take-downs, blocks, bans and censorship and stop jealously peering over the fence at the apparent successes of authoritarian regimes in stamping out speech and behaviour they do not like online. Instead, a liberal democratic approach to policing and online security must be introduced, ensuring security services are able to protect citizens while doing so in a way that is proportionate and with oversight.

Democratic state- and corporate-sponsored infrastructural growth is vital, including the export and promotion of infrastructure that bolsters the open web around the globe. Handing over the standards and roll-out of digital infrastructure to compromised providers and authoritarian regimes is unacceptable.

Case Study

Internet Shutdowns

Governments across the world are taking it as their sovereign right to take action against the open web: at the extreme through internet shutdowns, more-or-less sincerely to protect national security, law and order, or prevent online harms.[29, 30] However, shutdowns have been described as 'collective punishment'[31] for those affected, and impact not only fundamental freedoms of information and expression, but have significant negative economic and health effects.[32, 33] Internet restrictions lasting for months in Myanmar have been criticised in particular in 2020 for blocking access to essential information about the Covid-19 pandemic.[34]

Without clear global standards and commitments to what Internet access should be, we cannot determine when restrictions are legitimate or not. State-led attempts to deal with problems across platforms are leading to citizens' rights being eroded rather than protected, while also curtailing their ability to speak out in protest

Conclusions

It is a cliché to describe liberal democracy as a balancing act, but here we are again. Four forces will be responsible for the shape of the future Internet. State power, corporate power, individual power and the power of machines must be harnessed and managed in the name of liberal democracy. Correcting the balance of powers is the challenge we now face. The examples presented here show that moving too far in any one direction will undermine the project as a whole, and policy that ignores the importance of one of these powers will be insufficient.

There is evidence of failure wherever you look. The harms of the web in its current iterations are well-documented. We speak about the victories of the Internet less often these days, but this is a question of evidence too. In moving to a proactive vision of a liberal democratic Internet, we must celebrate and support the voices, designers and architects making positive strides forward, and ensure we hold all parts of the Internet to the standards of its success stories. There are lessons to be learned from Wikipedia, from StackOverflow, and from the legions of virtual communities that are thriving below the headlines.

There are also lessons to be learned from the web giants. They have rightfully come under fire over the past few years for their failings, but also contributed more than anyone to opening the Internet up to the world. Brought to the defence of liberal democracy, they may again be perceived as a vanguard of liberal democratic values around the world.

It is state and individual power where the most urgent questions must be answered. The internet will be the place where democracy is redefined in the 21st century, but doing so will require a radical improvement in state and multilateral governance of the online world and its underpinning technology. Ensuring individuals are able to exercise their rights online is a vital check on both state and corporate overreach.

The trilemma of states, individuals and the private sector is, however, not fit for the future. The accelerating development of machine power, from artificial intelligence to permissionless technology, will itself challenge all three for future dominance. Given the pace at which questions of global governance move, it is of crucial importance that steps taken reflect the growing influence of machines in our social, economic and political lives.

More than ever now, we need a vehicle to unite liberal democracies in advancing and advocating their own vision of the web. While authoritarian powers are increasingly coherent in promoting their vision, democracies are currently fractured, with fundamental differences in approach in North America, Europe and Asia. Yet there are underlying values and interests that unite us and must be articulated.

Without evidence for what works online, and without a principled vision for the internet, our democratic traditions, government and society will fall behind authoritarian states, industry giants and powerful technology in the race to reshape the most important international political, cultural and social space in existence. We must not focus on what we don't want, but rather verbalise what we do.

Footnotes

1. R. Larsen, The Political Nature of TCP/IP, Science, Technology & Society Program, University of Pennsylvania; https://repository.upenn.edu/cgi/viewcontent.cgi?article=1004&context=momentum (p.27)
2. Such as the battle between TCP and the Open Systems Interconnection protocol.
3. Larsen (p.47)
4. K O'Hara & W Hall, Four Internets: The Geopolitics of Digital Governance, CIGI Paper No. 206, December 2018; https://www.cigionline.org/publications/four-internets-geopolitics-digital-governance
5. Wikipedia, Technological singularity; https://en.wikipedia.org/wiki/Technological_singularity
6. Vint Cerf himself admitted the only thing preventing earlier adoption of encryption standards online was military security classification of the required technology; https://www.youtube.com/watch?v=17GtmwyvmWE&feature=share&t=23m1s
7. In countries where rule of law is itself compromised or alien, it is simply the state's ability to wield ultimate power.
8. The New York Times, How China Uses High-Tech Surveillance to Subdue Minorities, 22nd May 2019; https://www.nytimes.com/2019/05/22/world/asia/china-surveillance-xinjiang.html
9. S Dickson, China's New Cryptography Law: Still No Place to Hide, Harris Bricken, 7th November 2019; https://www.chinalawblog.com/2019/11/chinas-new-cryptography-law-still-no-place-to-hide.html
10. J L Zittrain, The Future of the Internet -- And How to Stop It, Yale University Press & Penguin UK, 2008; https://dash.harvard.edu/bitstream/handle/1/4455262/Zittrain_Future+of+the+Internet.pdf?sequence=1
11. A useful list of example DAOs can be found at https://hackernoon.com/what-is-a-dao-c7e84aa1bd69
12. G. De Gregorio, Democratising Online Content Moderation: A Constitutional Framework, Computer Law and Security Review, April 2020;
13. S. Zuboff, Big Other: Surveillance Capitalism and the Prospects of an Information Civilization (2015) p.83
14. B. Schneier, Data and Goliath (2015) p.58
15. Reuters, Coronavirus brings China's surveillance state out of the shadows, 7th February 2020; https://www.reuters.com/article/us-china-health-surveillance-idUSKBN2011HO
16. D Daniel & C I Speranza, The Role of Blockchain in Documenting Land Users' Rights: The Canonical Case of Farmers in the Vernacular Land Market, Frontiers in Blockchain, May 2020; https://www.frontiersin.org/articles/10.3389/fbloc.2020.00019/full
17. L Tombs, Could blockchain be the future of the property market?, HM Land Registry Blog, May 2019; https://hmlandregistry.blog.gov.uk/2019/05/24/could-blockchain-be-the-future-of-the-property-market/
18. European Commission, Ethics guidelines for trustworthy AI, April 2019; https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
19. The Guardian, China's hi-tech war on its Muslim minority, 11th April 2019; https://www.theguardian.com/news/2019/apr/11/china-hi-tech-war-on-muslim-minority-xinjiang-uighurs-surveillance-face-recognition
20. The Washington Post, Huawei worked on several surveillance systems promoted to identify ethnicity, documents show, 12th December 2020; https://www.washingtonpost.com/technology/2020/12/12/huawei-uighurs-identify/
21. The Atlantic, Defund Facial Recognition: I'm a second-generation Black activist, and I'm tired of being spied on by the police, 5th July 2020; https://www.theatlantic.com/technology/archive/2020/07/defund-facial-recognition/613771/
22. The Verge, Amazon bans police from using its facial recognition technology for the next year, 10th July 2020; https://www.theverge.com/2020/6/10/21287101/amazon-rekognition-facial-recognition-police-ban-one-year-ai-racial-bias
23. BBC News, Facial recognition use by South Wales Police ruled unlawful, 11th August 2020; https://www.bbc.co.uk/news/uk-wales-53734716
24. The Verge, ACLU sues facial recognition firm Clearview AI, calling it a 'nightmare scenario' for privacy, 28th May 2020; https://www.theverge.com/2020/5/28/21273388/aclu-clearview-ai-lawsuit-facial-recognition-database-illinois-biometric-laws
25. Council on Foreign Relations, Trump and Section 230: What to Know, 2nd December 2020; https://www.cfr.org/in-brief/trump-and-section-230-what-know
26. CNN, Trump says right-wing voices are being censored. The data says something else, 28th May 2020; https://edition.cnn.com/2020/05/28/media/trump-social-media-conservative-censorship/index.html
27. The Verge, Joe Biden wants to revoke Section 230, 17th January 2020; https://www.theverge.com/2020/1/17/21070403/joe-biden-president-election-section-230-communications-decency-act-revoke
28. Demos, The Great Cyber Surrender: How police and governments abandon cybercrime victims, November 2020; https://demos.co.uk/project/the-great-cyber-surrender-how-police-and-governments-abandon-cybercrime-victims/
29. Chatham House, Asia's Internet Shutdowns Threaten the Right to Digital Access, February 2020; https://www.chathamhouse.org/2020/02/asias-internet-shutdowns-threaten-right-digital-access
30. DW, India's internet shutdowns function like 'invisibility cloaks', November 2020; https://www.dw.com/en/indias-internet-shutdowns-function-like-invisibility-cloaks/a-55572554
31. Ibid; https://www.dw.com/en/indias-internet-shutdowns-function-like-invisibility-cloaks/a-55572554
32. Human Rights Watch, Shutting Down the Internet to Shut Up Critics, 2020; https://www.hrw.org/world-report/2020/country-chapters/global-5
33. Ibid; https://www.dw.com/en/indias-internet-shutdowns-function-like-invisibility-cloaks/a-55572554
34. The Wire, Human Rights Groups Criticise 'World's Longest Internet Shutdown' in Myanmar, June 2020; https://thewire.in/south-asia/myanmar-worlds-longest-internet-shutdown

ISD | Powering solutions to extremism and polarisation

Digital Policy Lab
Discussion Paper

Future Considerations for Online Regulation

About This Briefing

This paper outlines several key aspects of digital policy that will require further exploration, debate and consensus-building in the coming years. It highlights areas requiring additional clarity following the release of key policy proposals by the EU and UK in late 2020. The also explores the tensions and trade-offs that may arise from new regulation as internet governance increasingly comes under the auspices of states and regulators.

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Introduction

As outlined in the accompanying *Discussion Paper: National & International Models for Online Regulation*, the liability regimes that govern online platforms ('intermediaries') have barely changed since the 1990s or early 2000s in many contexts, including the US and EU. These regimes were based on the premise that freedom of speech and growth of the digital economy would be at risk if platforms were held directly responsible for user-generated content. Yet it has become clear that many companies, and social media in particular, have created business models that can pollute, distort or fragment the public sphere.

These business models have increasingly dictated the success or failure of certain content, and been shown to favour sensationalist or contentious posts, first to increase traffic and user engagement and in turn to maximise ad revenue. Over time, such a dynamic has contributed to an increase of hate, extremism, terrorism and disinformation online that is harming society and democracy writ large. Governments have pressured technology companies to take on more responsibility and be accountable for harmful activity on their platforms. However, the resulting informal, voluntary or industry-led responses have often reactive or issue-specific, and in many cases insufficient. Despite some improvements through such approaches, the overarching structures and processes that promote, amplify or recommend harmful content, or enable and encourage harmful activities, have typically remained opaque and unaffected.

As a result, governments are increasingly turning from informal, voluntary or industry-led self-regulatory efforts to legislation to tackle these negative externalities enabled by online platforms. In our previous paper, we broadly categorised this raft of recent and new regulations into two broad categories:

- **Content-based approaches**, often targeting a specific online harm such as hate speech or electoral disinformation, and focusing on the effective and timely removal of that content where appropriate (often referred to as a 'notice-and-takedown model').

- **Systemic approaches**, whereby online platforms must demonstrate that their policies, processes and systems are designed and implemented with respect to the potential negative outcomes that could occur, across a range of possible harms.

The relative lack of progress thus far in stemming the impact of online harms has led to greater support for the latter of these two options. Such systemic approaches aim to develop a unified oversight regime that can be used to address a range of illegal, and in some cases legal but harmful, activity; the intention being to impact the internal processes of the most dominant technology companies, who can play an important role in exacerbating them.

In December 2020, the European Commission presented its Digital Services Act (DSA) proposals fora new regulatory regime to encompass social media, online marketplaces, and other online platforms operating in the EU. The DSA represents a significant step towards a more systemic approach. It proposes a new set of obligations for digital services that act as intermediaries, with the intent to "create a safer online experience for citizens to freely express their ideas, communicate and shop online by reducing their exposure to illegal activities and dangerous goods and ensuring the protection of fundamental rights".[1] As such, it updates the e-Commerce Directive, in place since 2000, signalling a shift towards a more preventative (rather than reactive) model. Additionally, according to the European Commission, these new rules are "an important step in defending European values in the online space" and are intended to set "a benchmark for a regulatory approach to online intermediaries also at the global level".[2]

Also in December 2020, the UK Government released its latest proposals via a full consultation report to tackle a plethora of online harms, ranging from hate, extremism and terrorism to child safety, online abuse and disinformation. These are intended to form the basis of the upcoming Online Safety Bill, which seeks to "make the UK the safest place in the world to go online, and the best place to grow and start a digital business". This will be achieved via a new statutory 'duty of care' regime enforced by an independent regulator (now confirmed as the existing broadcast and telecommunications regulator, Ofcom).[3] These proposals are summarised in greater depth in the accompanying ***Policy Summary: EU Digital Services Act & UK Online Safety Bill*** paper.

Neither the EU or UK proposals are yet final, and will both be fiercely debated by numerous stakeholders within politics, industry and civil society in 2021. As such, both proposals will be subject to various amendments before they are finalised in the years ahead, and likely on an ongoing basis due to the rapid evolution of the online ecosystem.

Global Ambitions

Both the European Commission and the UK Government explicitly state a global ambition to help shape the direction of digital policy beyond their borders, and set future precedents for how key areas of the public internet should be governed and regulated on behalf of their citizens. At their core, both attempt to substantially alter the current (im)balance of power between democratically elected governments and institutions, and private companies, while also encouraging companies to take greater care in the design of their products and services.

These ambitions point to an ongoing geopolitical contest to define the future of internet governance, including the balance of power between governments (both democratic and authoritarian), private companies, citizens, and technology itself. We explore these issues in the accompanying paper, ***The Liberal Democratic Internet & The Good Web Project***, which outlines the pressing need to collectively define a principled vision for the internet. This vision must encapsulate key democratic values in the infrastructure and public spaces of the online world, in order to effectively compete with emerging authoritarian models of internet governance.

Given the extraordinary breadth of political systems and cultural traditions, the expectation should not be for common legislation and regulation across this diverse range of contexts. However, there is a need to collectively restate the common shared values and interests that do unite liberal democratic countries, and better apply these to the online world. It is therefore vital to move beyond a US- and Euro-centric focus on both harms and responses, and to arrive at a truly global, but adaptable, liberal democratic model for internet governance.

Looking Ahead: Key Questions Arising from Proposed Regulation

This paper explores some of the key challenges and tensions emerging from existing efforts to regulate across contexts, including from legislative developments in the EU, UK, Australia and beyond. Ongoing points of contention include:

- From Content to Systems: Liability & 'Safety by Design'
- Regulatory Jurisdiction – who sets and enforces the rules?
- Addressing 'Legal Harms'
- Regulatory Scope – who is being regulated?

Without criticising current efforts or implying that a simple set of solutions exist, the following sections explore the inherent tensions and trade-offs in these areas. They also identify potential future issues that may become more prominent over time as governments continue to regulate online platforms. We hope this can form the basis of future engagements via the DPL network, and add to the ongoing debate on digital policy.

From Content to Systems: Liability & 'Safety by Design'

Increasingly, proposals like those in the UK and EU are moving away from content-based notice-and-takedown approaches towards more systemic models of regulation. The former has led to some improvements in removing illegal content (e.g. terrorism), and encouraged companies to allocate more resources to tackling such problems. Existing policy initiatives in this area are reviewed in the accompanying *Discussion Paper: National & International Models for Online Regulation*. However, they have largely proved ineffective in stemming the tide of other online harms at scale (e.g. hate speech) or where they do not cross existing legal boundaries (e.g. disinformation). They also carry risks of encouraging overly broad content removal by platforms in order to avoid potential financial penalties (so-called 'overblocking').

Regimes that we label as 'systemic approaches' are those seeking to avoid such pre-emptive and overly-restrictive behaviour from platforms. Rather than fixating on individual instances of illegal or harmful content, a systemic approach seeks to protect legitimate free speech by incentivising proactive risk-prevention measures in the design of platform product, policies and processes. In the EU and the UK, proposals add an additional layer to existing liability provisions, preserving the underlying principle that platforms should not be held directly liable for user generated content. The measures outlined in the proposals intend to compel platforms to move towards 'safety by design' approaches that encourage proactive consideration of potential risks to their users, or negative societal externalities that could arise from the use of their products or services. They also contain various provisions to ensure that companies have consistent flagging, reporting and moderation systems in place, including appeals processes and improved transparency requirements. These would require companies to do a better job at explaining their decisions to those affected. Under both regimes, large platforms' systems would be subject to audit by independent regulators, including their core algorithms, to ensure they meet a sufficient threshold in harm prevention. These audits would detail the measures taken, data collected on efficacy and future steps for improvement.

These systemic approaches signal a major shift in the power dynamic between governments, technology companies and their users. The goal is to counterbalance existing commercial incentives with the need to better combat illegal content and activity and, in some cases, to reduce harms that can result from legal content. These systemic approaches should also start to address the structural nature of online harms via increased scrutiny and oversight on the underlying processes, such as recommendation and newsfeed algorithms, that can play a significant role in determining the reach and impact of content online.

Longer term, these new regulatory regimes should help to fuel competition, innovation and investment in online safety. Companies that are more effective and efficient at meeting regulatory demands will have a competitive advantage. Users may also increasingly seek out products and services that offer a safer online experience. While existing proposals take a graduated approach to regulation, with greater obligations and oversight for larger companies, it will be important to ensure that such measures do not permanently enshrine the current market positions of the internet giants. These larger companies, due to the scale of their user bases, undoubtedly face a more significant challenge, but also boast significantly more resources than many of their smaller counterparts offering similar products and services. Anti-trust and monopoly power will be addressed through other regulatory instruments, such as the EU's Digital Markets Act, but regulators will have an important role to play in sharing best practices where possible, and supporting smaller companies in meeting their new obligations. The private sector should also be encouraged to cooperate across industry and share resources and expertise.

The introduction of new regulation should make a significant impact on the 'supply-side' of online harms, and reduce the levels of illegal or harmful content found online. However, truly systemic approaches must also include provisions considering the 'demand-side', for example through greater whole-of-society investments in digital and media literacy education. Existing initiatives that recognise this include the EU's European Democracy Action Plan, a partner proposal to the DSA, and the emerging Media Literacy Strategy in the UK. New online regulation may be necessary, but will not necessarily be sufficient in tackling the full range of online harms these proposals are intended to address.

Designing Meaningful Transparency

With both voluntary industry responses and legislation, the importance of transparency around the policies, processes and outcomes of platform actions has become ever clearer. As argued in the accompanying *Discussion Paper: Transparency, Data Access and Online Harms*, transparency should enhance public understanding of online harms and which policy responses are most effective, and contribute to building stronger norms around what is acceptable online. There are difficult questions to address in setting expectations and rules for the types of data and levels of access that governments should require of different companies, in order to meaningfully assess their success or failure in meeting regulatory standards.

Under emerging proposals, enhanced transparency requirements would provide vital data for regulators and researchers on the impact of company measures to tackle illegal content. They also seek to provide additional clarity for those whose content is (correctly or incorrectly) removed. However, it remains to be seen how effective these demands for transparency on company policies and actions will be: to a certain extent, the suggested requirements re-emphasise a reliance on private sector transparency reporting processes over external democratic or judicial scrutiny. Affected users usually have to exhaust appeals through companies' systems before being able to seek truly independent arbitration and redress. There will remain a tension between working within existing approaches to transparency (forged mostly by the companies themselves) and seeking to invent new types of external audit on the actions, processes and outcomes of company actions. This tension presents a choice between a) practical but potentially limited regulatory regimes that work to improve or amend existing private sector approaches to transparency and b) ambitious efforts to create entirely new types of information access requirement and assessment for these types of challenges.

Regulatory Jurisdiction – who sets and enforces the rules?

Both the EU and UK express ambitions to help set the direction of internet regulation globally, and their respective proposals broach the challenge of applying national-level or regional jurisdiction to inherently global internet services. Both approaches raise questions as to the potentially enormous scope, complexity and viability of enforcement by the relevant regulators. The challenge of how to build national or regional regulation for companies whose user base often lacks identifiable locations, that can play host to transnational crimes, and that base their operations in just one or two countries but function globally, remains critical.

Terrorist and extremist content, hate speech and disinformation are all inherently global phenomena. Islamist, far-right and state and non-state disinformation networks are also increasingly transnational. This means that a significant proportion of such content consumed by national audiences will originate outside of these jurisdictions. Such forms of criminal activity or harm are also characterised by a longstanding inconsistency (and abuse in certain contexts) of international definitions, including significant variance in other legal instruments such as terrorism designation or proscription lists.

In terms of **scope**, large proportions of content on social media are available to any user of the service (and in many cases any internet user regardless of whether they have an account on that service, e.g. YouTube). This is true regardless of the origin of the content or the location of the ultimate content consumer, and raises questions as to the feasibility of enforcing regulation around nationally defined threats. For example, given the vast diversity within the EU[4] and the UK[5] populations, audiences are likely to consume content in several hundred languages, which will inevitably include illegal content such as hate speech, as well as legal but harmful content such as disinformation. **While there is some precedent for regulators covering a wide array of content and languages (e.g. Ofcom regulates satellite television providers in the UK), the enormous scale of online content represents a significant change in the order of magnitude.**

Case Study

The DSA's efforts to streamline approaches across the EU

- The proposals apply to all intermediaries that provide their services in the EU regardless of where they are primarily legally established.
- This encompasses operations based in the EU, platforms with a significant number of EU users, and services targeted towards EU Member States (for example using a national language, currency or domain such as .fr, .de etc.).
- Under the DSA, every company will need to assign a legal representative based within the region, who can be held accountable for non-compliance or other issues as they arise.
- Enforcement and oversight will fall to the Digital Services Coordinator (DSC) in the relevant country, who will present and/or escalate issues to the EU Board as needed. Where companies fail to take necessary steps (e.g. following an independent audit by the DSC), EU-level pressure can also be applied, including the levying of substantial fines based on annual turnover.

Case Study

The UK Online Harms response

- The UK Government response of December 2020 suggests that companies will fall under the scope of the upcoming Online Safety Bill if they host user-generated content available to UK users, or facilitate online interaction (either public or private) where one or more participants are based in the UK.

The Online Harms proposals still seek to apply UK laws to online content originating elsewhere, whenever that content is available to UK users. The DSA, as horizontal EU legislation, will rely on a complex **patchwork of national laws** in EU Member States with varying legal systems and traditions. While there is typically widespread agreement that what is illegal offline should also be illegal online, national variations within the EU could create new tensions given the inherently transnational nature of the internet. For example, some EU Member States have banned certain terrorist or violent extremist groups, but these are not banned in other national contexts. Alternatively, content denying the Holocaust or praising the Nazi regime are illegal in Germany, but not elsewhere within the EU. In practice, this is likely to result in companies including all categories of content that are illegal somewhere into their Terms of Service that apply everywhere, to avoid the added complexity of differential rules applying in different markets.

While this may already be the case in certain areas that are less likely to raise serious objections (e.g. Facebook explicitly prohibited Holocaust denial under its TOS for the first time in 2020)[6], overall this approach could result in the de facto extrapolation of national laws beyond national borders. This could prove problematic within the EU, for example as certain Member States' governments hold divergent views on LGBTQ+ or gender equality; it will also create additional pressure on companies to adhere to content-focused laws via their TOS, even in contexts that have laws at odds with core EU values (e.g. blasphemy laws).

The potential implications of this could be significant. One outcome may be that platforms geo-fence their services to exclude countries or regions that require adherence to regulation in order to avoid compliance with new rules, as with the emerging standoff in the Australian case.[7] Alternatively, they may be over-cautious and limit content everywhere via their TOS based on the most stringent national laws. While this might lead to some improvements in safety, it would also result in a loss of access and choice for users, and potentially harm competition and freedom of speech.

On **enforcement**, emerging proposals include examples of provisions around liability for senior leadership of services; e.g. the requirement to designate a point of contact and a legal representative who could be held liable for non-compliance. However, this approach raises questions as to whether these could be effectively applied and enforced in cases of non-compliance where the individuals in question do not reside permanently in the EU or UK, or could lead to services withdrawing from either jurisdiction to avoid regulation.

The benefits and challenges of regional-level policymaking are well demonstrated by the difficult enforcement environment that the DSA faces in the EU. The approach taken will place considerable responsibility for oversight and enforcement in the hands of particular national regulators where platforms are based within the EU (e.g. many of the largest companies are established in the Republic of Ireland). This could create a disproportionate imbalance of power across different national regulators, with harms experienced in one country becoming the responsibility of the regulator in another, but who would not be accountable to that national government. Thinking beyond Europe, these jurisdictional tensions will continue to challenge any efforts to design supra-national regulatory structures, despite the many potential advantages that come with harmonisation of approaches across national borders.

Addressing 'Legal Harms'

Governments are rightly concerned about mandating specific action against so-called legal harms; heavy-handed laws may stifle innovation, exclude marginal voices or negatively impact fundamental rights to freedom of speech. Achieving a balance between protecting expression and ensuring a fair and safe field for speech while not exacerbating risks of physical or psychological harm depends on the ability to clearly define what 'harm' means. The protection of fundamental rights, including those relating to security, safety and free expression, must be supported by a proportionate and evidence-based understanding of the boundaries for a platform's responsibilities. Approaches to regulation that attempt to incorporate and therefore to define legal but 'harmful' content or activity must answer a number of key questions. Those questions include:

- **Measuring harm**: How will causality be established between exposure to online content or activity and physical or psychological harm to an individual or individuals? What offline and online evidence might be helpful in these types of assessment? What can we learn from other areas of regulation that deal with the externalities of businesses or corporations? How can governments design thresholds to determine adverse psychological impacts from online content or activity?

- **Determining scope of harm**: Should social harms – those that threaten the institutions, practices or processes of democracy, for example – be included alongside individual harms in the scope of regulation? What are the additional challenges in identifying, measuring and therefore preventing or sanctioning these types of societal harm?

Inevitably, the broader the definition applied, the higher the risk of a disproportionate impact on freedom of speech. If a broad definition is used, then the regulator would have more scope to determine what should be considered harmful. This would create a greater incentive for platforms and services to cast a wide net when considering the types of content or activity that would be in scope, and help them to avoid further regulatory scrutiny or enforcement.[8] **Conversely, if no attempts are made to delineate legal versus illegal speech, it places undue pressure on platforms' Terms of Service to arbitrate public discourse without any level of oversight from independent institutions**. Almost inevitably, the attempt to further define the boundaries of legal but harmful speech will result in significant grey areas, as seen with existing debates over platforms' Terms of Service that outline legal content violating their own standards.

It is worth considering how regulatory efforts that avoid including legal but harmful content in their remit might still usefully affect the presence or scale of these types of harm online. Systemic approaches to regulating illegal content on digital platforms can include robust requirements to improve transparency, incorporate safety-by-design principles and practices, and allow independent auditing powers to oversee the policies, processes and outcomes of company actions. In those cases, platforms may well have to adjust and improve all processes relating to the transparency and safety of content moderation and curation on their services. **While these steps would be taken by platforms to abide by their responsibilities regarding illegal content, the transparency improvements may also limit the presence of 'legal' harms – or at least improve public understanding of how and where those harms materialise**.

Case Study

The UK's Online Harms response places obligations on Category 1 platforms for tackling legal but harmful content:

- All platforms will have an obligation to tackle illegal content (e.g. terrorism, CSEA), and consider the risks posed to children by legal but harmful content, but Category 1 companies will be required to address legal but harmful content for all users in their Terms of Service (TOS) and demonstrate that they enforce them consistently through transparency reporting.

- Whilst the UK Government has published draft Codes of Practice for terrorism and CSEA content, which will set out a range of suggested (but largely non-binding) approaches for tackling these categories of illegal content, it will fall to the regulator to establish codes for the legal but harmful category.

- The UK proposals are fundamentally based around the concept of 'harm', defined as when online content or activity "gives rise to a reasonably foreseeable risk of a significant adverse physical or psychological impact on individuals". To provide additional clarity, the forthcoming legislation will further define the types of harmful content and activity in scope, accompanied by several priority categories of harm included in secondary legislation, and some specific exceptions where regulation is already in place (e.g. IP rights, data and consumer protection, cyber-fraud, hacking).

Case Study

The DSA does not attempt to explicitly regulate legal but harmful content, and proposes a series of existing or additional legislation to cover illegal content, which is otherwise not defined, such as existing Union or Member State laws, or the proposed Regulation on Terrorist Content Online:

- The DSA maintains existing liability rules for providers of intermediary services, whereby a hosting service is only obliged to act when it obtains actual knowledge or awareness of illegality – at that point it must act in a timely manner to remove or disable access to that content.[9]

- The "harmful information and activities" category is described as "a delicate area with severe implications for the protection of freedom of expression".[10] However, the DSA proposals do acknowledge that the scale and ubiquity of certain platforms has "increased their role in the intermediation and spread of unlawful or otherwise harmful information and activities".[11]

- As a result, the proposals would create due diligence obligations that cover platforms' content moderation for both illegal content and legal but potentially harmful content that contravenes their TOS, but stresses that the two categories should treated differently and rules mandating the removal of legal content would not be included.[12]

- The DSA proposals highlight the broader risks posed by the largest platforms that are "optimised to benefit their often advertising-driven business models [...] without effectively identifying and mitigating the

risks and the societal and economic harm they can cause".[13] As a result, these platforms will be required to conduct assessments that cover any systemic risks related to their services, including potential misuses by users, and then take appropriate mitigating steps.

- This will cover both illegal content and activities, but also other negative impacts on fundamental rights such as respect for private and family life, freedom of expression and information, the prohibition of discrimination, and the rights of children. It also includes "the intentional and, oftentimes, coordinated manipulation of the platform's service, with a foreseeable impact on health, civic discourse, electoral processes, public security and protection of minors, having regard to the need to safeguard public order, protect privacy and fight fraudulent and deceptive commercial practices".[14]

Finally, the Coronavirus pandemic has demonstrated the potential for extraordinary circumstances to arise in which platforms can become the services through which new types of harmful or threatening content and activity emerge. A number of existing proposals for digital regulation make reference to 'crisis protocols' that could be activated in these types of extraordinary circumstance, including events that affect public security or health. Examples might include a natural disaster, pandemic, act of terrorism or major election or, in the words of the DSA, cases where "online platforms may be misused for the rapid spread of illegal content or disinformation or where the need arises for rapid dissemination of reliable information."[15] The exact thresholds for 'crisis' and expected response remain unclear, but could be defined in partnership with civil society and other expert bodies. Given the current lack of precision in predicting what such emergencies might be, the related responsibilities for platforms in these circumstances are likely to remain light-touch and rely on interpretation of, enforcement of and transparency around their own Terms of Service.

Regulatory Scope – who is being regulated?

A key tension in designing effective, but also efficient and proportional, regulation for online platforms is defining what types of service and/or company fall under specific obligations. A central contention in emerging policy proposals in this area is whether 'harm' should be defined by reach (i.e. dangerous content reaching the widest audience) or extremity (i.e. how dangerous that content is).

If the objective is to regulate for online safety, then the number of platform users is not necessarily an accurate proxy for harm. Recent years have seen a series of migrations from larger platforms, which have gradually improved moderation and enforcement under pressure from governments (at least in certain areas e.g. terrorist content), towards smaller platforms that are either ill-equipped or unwilling to take action. As further regulations are introduced, predominantly focused on the largest platforms, this trend is only likely to continue, for example the recent growth in users seen on platforms such as Parler and Telegram.

Such platforms present a different safety challenge depending on the type of harm. Extremist groups may start on larger platforms, but radicalisation and recruitment often take place in smaller, more ideologically homogenous and unmoderated spaces, where it can be more effective. In contrast, disinformation campaigns can be organised on smaller platforms by non-state actors, but still require exposure to the broader audiences that larger platforms provide to achieve maximum impact.

While the size of a platform is by no means a perfect analogue for digital services, it is encouraging that emerging proposals in liberal democracies are beginning to take into account the wide variety of different services provided by platforms, and at different levels of the internet's technical 'stack' - for example by distinguishing between infrastructure providers and platforms. While imperfect, this is certainly preferable to a blanket approach that applies regulation designed with the most dominant companies in mind across the board.

Both the EU and UK proposals recognise that not all internet services present the same scale of challenges in terms of online safety, or have comparable levels of internal resources and expertise to address them. Additionally, both stress the importance of not dis-incentivising growth and innovation, and creating disproportionate regulatory burdens on smaller companies or start-ups. As such, both proposals include a graduated approach to regulation, with larger platforms facing additional obligations and oversight in comparison to their smaller counterparts. These are explored in greater depth in the accompanying Policy Summary paper.

Over time, the enhanced transparency provisions in both proposals will provide regulators and researchers with greater access to data, and therefore foster an improved understanding of the role online services play in facilitating online harms. Once this is achieved, it may be possible to move away from scale as the key determinant of risk, and towards a more nuanced approach to determining the responsibilities and obligations placed on online services.

Looking Ahead

This paper has only scratched the surface of challenges that continue to emerge as governments grapple with the most effective, proportionate and achievable ways to design regulation for the digital environment. None of these has a simple answer that will apply to each and every context in which regulation is being considered. Complicated webs of regulation look set to dramatically alter the way in which internet services can and must serve their users over the coming decade. Beyond the questions of scope, approach, definitions and jurisdictions set out above, governments face the additional puzzle of how to fit such regulation alongside parallel efforts in the realms of competition and data privacy. The former includes questions of antitrust action, but also potential issues around existing trade secret legislation which may limit transparency of internal company processes. The latter is particularly pertinent to the ongoing policy debates of how to regulate encrypted messaging platforms while maintaining users' rights to private communication.

Legal, political and cultural contexts will need to shape the solutions that each government crafts if regulation is to serve the public as effectively as possible. But these types of regulatory challenge are not entirely new: complex oversight mechanisms have been designed in countless areas of corporate business and political activity that each contain seemingly intractable tensions between rights, responsibilities and risk. A multidisciplinary conversation between liberal democratic governments can help guide decision-making on these issues, but must take lessons from across sectors and across borders into account.

Endnotes

1. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348
2. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348
3. Online Harms White Paper: Full government response to the consultation. https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response#part-1-who-will-the-new-regulatory-framework-apply-to
4. 24 official languages, and over 60 regional languages or dialects, e.g. Catalan or Basque: https://europa.eu/european-union/about-eu/eu-languages_en
5. Alongside English, Welsh and Gaelic, the 2011 UK Census lists Polish, Panjabi, Urdu, Bengali, Gujarati, Arabic, French, Chinese (all dialects), Portuguese and Spanish as the next most common languages. The British Council estimates over 300 languages are spoken in London alone: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language & https://study-uk.britishcouncil.org/moving-uk/student-life/language
6. https://www.bbc.co.uk/news/technology-54509975
7. https://www.bbc.com/news/world-australia-56107028
8. This is essentially the danger of 'overblocking' that has been at the heart of much of the criticism directed against content-based regulatory approaches and particularly the German Network Enforcement Act.
9. DSA (Article 5.1, §22)
10. DSA (p.9)
11. DSA (§5)
12. DSA Article 2.p
13. DSA (§56)
14. DSA (Article 26, §57)
15. DSA (§71)

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org