TRANSATLANTIC WORKING GROUP

A Cycle of Censorship:

The UK White Paper on Online Harms and the Dangers of Regulating Disinformation†

Peter Pomerantsev, Senior Fellow, Institute of Global Affairs,
London School of Economics and Political Science[1]

October 1, 2019

Introduction

This document contains two parts. The first is a summary of the UK Government Online Harms White Paper, including an overview of the arguments around it, responses to it and associated proposals by UK organisations in this field.

The second part proposes a way for the UK Government to reframe the challenge of "disinformation" as currently formulated in the White Paper. It argues that the current approach to combatting "disinformation" risks reinforcing a cycle of censorship worldwide – when the UK's publicly avowed role is to support freedom of expression globally. Another way forward is possible, one that strengthens the UK's commitments to upholding democratic values and human rights while combatting online deception.

The views and opinions expressed in this article are solely those of the author.

Contents

---

† One in a series: A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Part 1: White Paper Summary and Overview of Responses

After a year of heightened discussion in the UK about the dangers of hostile state online interference in UK democratic processes, non-transparent political campaigns by local actors, harassment of MPs and bullying of children as well as the persistent presence of violent extremism, the UK Government published an Online Harms White Paper[2] in March 2019. The White Paper attempts to find a middle way between self or no regulation on the one hand, and the imposition of liabilities for every piece of content on the other. It proposes a model of regulation that focuses on mitigating what it calls "harms" and putting systems in place to minimize "risk." So instead of being liable for each piece of content, companies are responsible for maintaining mechanisms to protect people and "society." Tech companies will be expected to follow codes of conduct and statutory "duties of care," enforced by an independent regulator.

The White Paper is a self-conscious attempt to approach internet regulation in an innovative way, and for the UK to take a leading global role on this issue. It is refreshing in its focus on the architecture of the internet, though it has also left many questions unanswered about what exactly an "online harm" is, let alone how one would mitigate it. The White Paper has had a striking number of responses in mainstream media as well as among the expert community, with a lively exchange of ideas between a variety of critics and defenders. It has shown structural differences between freedom of expression and human rights advocates on the one hand, and politicians' desires to reform the online space. The lack of consensus is worrying as it betrays fundamental lack of clarity about the UK's vision for the future of the information space, clearly a critical area for democracy.

The analysis below first addresses the main ideas in the White Paper and the overall criticisms of them. While the White Paper is the main focus, I also look at studies and proposals that have influenced the debate in the UK:

- The Digital, Culture, Media and Sport Committee's final report on Disinformation and 'fake news'[3]
- The UK broadcasting regulator, OFCOM, report on Addressing Harmful Online Content[4]
- Carnegie UK's proposal for Internet Harm Reduction[5]
- Full Fact (UK's leading fact checker) paper on Tackling Misinformation in an Open Society[6]

Summary of White Paper Proposals

Very broad scope: The Online Harms White Paper foresees creating mechanisms to regulate companies "that allow users to share or discover user-generated content or interact with each other online," e.g., "two main types of online activity": hosting, sharing and discovery of user-generated content, and facilitation of public and private online interaction between service users. No exception is allowed based on size or nature of the services. However, the framework should ensure a differentiated approach for private communication, "meaning any requirements to scan or monitor content for tightly defined categories of illegal content will not apply to private channels." The regulator will provide support to start-ups and small- to medium-sized enterprises (SMEs) "to help them fulfill their legal obligations in a proportionate and effective manner."

A new statutory duty of care will be introduced to "make companies take reasonable steps to keep their users safe and tackle illegal and harmful activity on their services." These will be fleshed out with specific codes of practice targeted to each specific type of harm (see below). However, companies will still need "to be compliant with the overarching duty of care even where a specific code does not exist," continuing to assess and respond "to the risk associated with emerging harms or technology."

Online harms: the codes of practice will cover harms "**with clear definition,"** such as child sexual exploitation, terrorist content and activity, organised immigration crime, modern slavery, extreme pornography, revenge pornography, harassment and cyberstalking, hate crime, encouraging or assisting suicide, incitement of violence, sale of illegal goods/services (such as drugs and weapons on the open internet), content illegally uploaded from prisons, or sexting of indecent images by under 18-year-olds (creating, possessing, copying or distributing indecent or sexual images of children and young people under the age of 18). They will also cover harms "**with a less clear definition,"** such as cyberbullying and trolling, extremist content and activity, coercive behaviour, intimidation, disinformation, violent content, advocacy of self-harm, and promotion of female genital mutilation. **Underage exposure to legal content** will also be in scope, such as children accessing pornography, children accessing inappropriate material (including under 13 years of age using social media and under 18 years of age using dating apps; and excessive screen time).

Online harms excluded: all harms to organisations (e.g., competition law, most cases of IP violation) and all harms suffered by individuals resulting directly from a breach of the data protection legislation and from a breach of cyber security or hacking are outside the scope of paper, as these are covered by other applicable laws.[7]

Codes of practice will set out how this duty of care should be fulfilled (through "systems, procedures, technologies and investment, including in staffing, training, and support of human moderators"). Companies that do not wish to comply with these codes will have to explain how what they will be doing will have the same or a greater impact. These codes of practice will be drafted:

(1) under the direction of an independent (existing or to be created) regulator;

(2) under the direction of the British government for codes of practice on terrorist activity or child sexual exploitation online; and

(3) in cooperation with **law enforcement for "illegal harms"** (such as incitement of violence and sale of illegal goods and services).

Compliance with this duty of care will be overseen and enforced by the regulator. This regulator will "set out expectations for companies to do what is reasonably practicable to counter harmful activity or content, depending on the nature of the harm, the risk of the harm occurring on their services and the resources and tech available to them." When assessing compliance, the approach taken will be "risk-based," "prioritising regulatory action to tackle harms that have the greatest impact on individuals or wider society" and will consider "whether the harm was foreseeable, and therefore what is reasonable to expect a company to have done."

Upload filters: When overseeing the implementation of the codes of practice, the regulator "will not compel companies to undertake general monitoring of all communications on their online services." However, the

British government considers that "there is a strong case for mandating specific monitoring that targets where there is a threat to national security or the physical safety of children."

Enforcement power: The regulator will be able to issue substantial fines, to ask for the publication of annual transparency reports, to require additional information from companies (including the use of algorithms in selecting content for users), to oversee the implementation of user redress mechanisms, to promote the development and adoption of safety technologies to tackle online harms and to "encourage access of independent researchers" to the data of companies. Additional powers to **disrupt the business activities** of a non-compliant company may be granted following the consultation, e.g., measures forcing third-party companies to withdraw any service they provide that directly or indirectly facilitates access to the services of the non-compliant company, ISP blocking and/or imposing **liability on individual members of senior management**. The regulator will be funded by industry in the medium term, and will have a "legal duty to pay due regard to innovation, to protect users' rights online."

White Paper and 'Viral Deception'

The White Paper claims the regulator will not "police truth on the internet" but mitigate the harms caused by "disinformation" and "online manipulation," which constitute "Threats to Our Way of Life":

> Our society is built on confidence in public institutions, trust in electoral processes, a robust, lively and plural media, and hard-won democratic freedoms that allow different voices, views and opinions to freely and peacefully contribute to public discourse.
>
> Inaccurate information, regardless of intent, can be harmful – for example the spread of inaccurate anti-vaccination messaging online poses a risk to public health. The government is particularly worried about disinformation (information which is created or disseminated with the deliberate intent to mislead; this could be to cause harm, or for personal, political or financial gain).
>
> Disinformation threatens these values and principles, and can threaten public safety, undermine national security, fracture community cohesion and reduce trust.

This is very vague language, but as the White Paper elaborates on "disinformation" and "online manipulation" it makes a laudable move away from obsessing over content to focus on behavior, actors and online architecture, though in a somewhat unstructured way.

Under "threats" from online disinformation, the White Paper quotes studies from the Oxford Computational Propaganda Project about the extent of organised social media manipulation, highlights the risks posed by micro-targeting and identifies the disinformation campaigns of the Russian state as particularly noteworthy. The White Paper describes the "impact" of this threat as being that people are largely not aware that algorithms define what they see online, and that their personal data, browsing history and networks play a part in this. Only 3 in 10 adults, the White Paper states, are aware of how companies collect people's data online.

"Online manipulation" seems a slightly wider category of harm than "disinformation." The White Paper argues that while "tolerance of conflicting views and ideas are core facets of our democracy" they are "inherently vulnerable to the efforts of a few to manipulate and confuse the information environment for nefarious purposes, including undermining trust. A combination of personal data collection, AI based

algorithms and false or misleading information could be used to manipulate the public with unprecedented effectiveness."

The White Paper makes a distinction between "legitimate influence" and "illegitimate manipulation." Examples of the latter in broadcasting regulation include subliminal messaging and other techniques which influence people without them being aware. Again, a recurring motif seems to be that people's lack of understanding about how and why the information environment around them is being shaped in certain ways is in itself a harm.

When it comes to fulfilling the Duty of Care on disinformation, the White Paper suggests that "companies will need to take proportionate and proactive measures to help users understand the nature and reliability of the information they are receiving, to minimise the spread of misleading and harmful disinformation and to increase the accessibility of trustworthy and varied news content."

The areas the regulator will include in a code of practice can be broken down into several categories.

i) Transparency: measures to ensure people understand what they are seeing online and why. People should know "when they are dealing with automated accounts," and there needs to be more clarity around political advertising.

This begs the question of what exactly is "political advertising." Electoral law defines electoral material as "material which can reasonably be regarded as intended to promote or procure electoral success at any relevant election." As Full Fact argue, this is too narrow a definition, and there is a need for transparency around political messaging outside of campaigns. Full Fact believe that advertising transparency requires full information on content, targeting reach and spend, which must all be provided in real time. They call for all factual claims used in political adverts to be pre-cleared, and compulsory watermarks to show the origin of online adverts.

ii) Cooperation with Fact Checkers and Boosting Authoritative News: The White Paper says companies will need to make "content which has been disputed by reputable fact-checking services less visible to users" and promote "authoritative news sources." Companies will need to promote "diverse news content, countering the 'echo chamber' in which people are only exposed to information which reinforces their existing views." It quotes the Cairncross Review, which "proposed that a 'news quality obligation' be imposed upon social media companies, which would require these companies to improve how their users understand the origin of a news article and the trustworthiness of its source."

These exhortations are, however, difficult to enact in practice. Fact-checking bodies have had a mixed experience working with technology companies. There is a problem due to the difference in scale between tech companies and fact-checkers, with fact-checkers feeling that they have no impact on how content is eventually shown online and fear they are being used as PR cover; there are also differences of opinion about which sorts of dis-, mis- and mal-information to flag.

Defining "quality news" is also difficult. Two approaches predominate: Voluntary associations that a media organisation can join, membership of which guarantees certain standards. This, however, risks leaving out the credible sites (blogs, Facebook publications in authoritarian countries, etc.) that have not joined such associations. Another approach is AI driven, and tries to automatically check for factual errors on domains or

suspicious metadata that suggests sites have been put together too quickly to be professional. Such AI-driven approaches are still highly unreliable.

iii) Algorithmic Accountability: Point 7.30 in the White Paper demolishes the current business model of at least one well-known tech company: "Companies will be required to ensure that algorithms selecting content do not skew towards extreme and unreliable material in the pursuit of sustained user engagement."[8]

The line has no further elaboration in the White Paper, but it could be of vast significance. The White Paper provides no clarity how any algorithmic oversight would work in practice. Will it entail being given access to the process of how tech companies create and adjust their algorithms? In which case how will competitive advantages and innovations be preserved? Or will there be some method of judging whether technology companies' promises to adjust algorithms are being followed?

Responses to the White Paper

The White Paper has drawn sharp criticism, which roughly falls into two camps.

The first line of criticism stresses the "chilling effect" on free speech, fearing that companies will be over-zealous in their takedowns given how tough the potential sanctions are. Index on Censorship, for example

> is particularly concerned about the duty of care. The concept is closely linked to the 'precautionary principle,' which has been widely applied in the environmental field, where it means not waiting for full scientific certainty before taking action to prevent harm. This makes sense. However, applying the precautionary principle to freedom of expression runs a high risk of legitimising censorship, especially when combined with large fines. It creates a strong incentive for online platforms to restrict and remove content.[9]

Article 19 "strongly opposes any 'duty of care' being imposed on Internet platforms. We believe a duty of care would inevitably require them to proactively monitor their networks and take a restrictive approach to content removal."[10]

In her Twitter feed Professor Lilian Edwards points out the threats to block information service providers to a country could already be in breach of ECHR rules (ECHR Article 10) and the e-Commerce Directive stating that information service providers will not be made subject to prior authorisation "or other requirements having equivalent effect."

Another line of criticism focuses on the nebulous definition of "harm" in the White Paper. Confusingly the White Paper says harms will be "evidence based," but provides no evidence of harm. The well-known right-wing columnist Toby Young[11] complains "the word 'harm' isn't defined, even though it appears in the title. That's alarming because the white paper says the new regulator will ban online material 'that may directly or indirectly cause harm' even if the content in question is 'not necessarily illegal'. As an example of what it has in mind, the government singles out 'offensive material', as if giving offence is itself a type of harm. Merely showing that it hasn't caused the complainant any tangible harm won't be sufficient, since all the regulator will need to show is that it *may* cause them *indirect* harm." If we already have legislation on illegal harms (such as child sexual exploitation), this argument runs, why do we need more?

Many critiques stress the difficulties of transferring offline to online harms. Index on Censorship argue that "although social media has often been compared to the public square, the duty of care model is not an exact fit because this would introduce regulation – and restriction – of speech between individuals based on criteria that is far broader than current law." Graham Smith, a lawyer specializing in internet law and author of the Cyberleagle[12] blog, makes the point that while it is tempting to draw a simple comparison between offline and online "duties of care," it is worth bearing in mind that the former:

- are restricted to objectively ascertainable injury;
- rarely impose liability for what visitors do to each other; and
- do not impose liability for what visitors say to each other.

Smith argues that online "harms" and "duties of care" could overstep these boundaries. Smith goes on to say that the proposed online duty of care is no duty of care at all. A "proper" duty of care, he argues, is a "legal duty owed to identifiable persons. They can claim damages if they suffer injury caused by a breach of the duty. … Occasionally a statute creates something that it calls a duty of care, but which in reality describes a duty owed to no-one in particular, breach of which is (for instance) a criminal offence." Smith quotes an environmental law in respect of waste disposal, but which is nevertheless precise "about the conduct that is in scope for the duty." He concludes that this is a mechanism "unsuited to what people say and do online."

The White Paper leaves it to the regulator to decide what exactly "legal but harmful" behavior and activity entails in practice, and what precisely would be demanded of tech companies to show they have provided "duty of care": "its very *raison d'etre* is flexibility, discretionary power and nimbleness," writes Smith. "Those are a vice, not a virtue, where the rule of law is concerned."

For supporters of regulation, one response to such criticism has been to reference British broadcasting regulation. The DCMS Parliamentary Committee on "Fake News," for example, argues that the regime for regulating broadcast content standards could be used "as a basis for setting standards for online content." The Carnegie UK paper on Internet Harm Reduction[13] quotes a House of Lords debate on a social media duty of care where Baroness Grender argued that competent regulators have had little difficulty in working out what harm means:

> "If in 2003 there was general acceptance relating to content of programmes for television and radio, protecting the public from offensive and harmful material, why have those definitions changed, or what makes them undeliverable now? Why did we understand what we meant by "harm" in 2003 but appear to ask what it is today?"

OFCOM's task in the Communications Act 2003 to which Baroness Grender refers is somewhat harder than merely harm:

> "generally accepted standards are applied to the content of television and radio services so as to provide adequate protection for members of the public from the inclusion in such services of offensive and harmful material."

OFCOM's own paper on the subject, however, is more circumspect. It highlights the difference between broadcasting and online content. Not only are volumes of content much greater online, but the public also

could have very different expectations for online content. The multinational nature of platform operators is an added complication.

The OFCOM paper argues that the priorities for online standards setting should be in protecting minors, protection from illegal content, bullying and "trolling." Setting standards in "news," however, is complex.

The fact that the government regulator is a priori at odds with the government's position on regulating political "disinformation" and "fake news" is indicative of the systemic disagreements on how to tackle this area.

Conclusion

Though an innovative approach to regulating the online space, and one that does well to go beyond ill-advised attempts to regulate all user-generated content, the White Paper has several structural faults that will need to be overcome in future drafts.

Among those are two that are especially critical for the government to address. It will need to:

- articulate more clearly how it plans to defend freedom of expression online; and
- draw clear distinctions between the regulatory framework for illegal content on the one hand, and on what it terms "harmful but legal" content on the other. Two such vastly differing categories cannot reasonably be placed under one framework.

Due to the amount of criticism directed at the White Paper, a revised version is expected at the start of 2020.

Part 2: Breaking the Cycle of Censorship

The dangers of regulating disinformation, and how the UK Government can reframe the debate to undermine dictators and strengthen democracy

In July 2019, the UK Government held the Global Conference on Media Freedom. Hosted by the British Foreign Minister, the government claimed the conference to be the "first of its kind,"[14] "part of an international campaign to shine a global spotlight on media freedom and increase the cost to those that are attempting to restrict it." Human Rights lawyer Amal Clooney gave the opening speech in front of an audience of journalists and activists from across the world, many of whom have faced vicious attacks and censorship from governments in their home countries. Clooney spoke movingly about her work as a human rights lawyer defending journalists in oppressive regimes and about how freedom of expression is in danger.[15] The conference was a high-profile statement of intent from the UK, as it plans to make media freedom one of its headline foreign policy priorities.

In the same month that the British government hosted this highly publicized conference, freedom of expression organisations were submitting their response to the government's "White Paper on Online Harms." The White Paper proposes to impose a mandatory "duty of care" on tech companies that will force them to show they are mitigating both clearly illegal activity as well as what the government terms "legal but harmful" content on their platforms, including disinformation. In their responses to the White Paper, freedom of expression groups cast the UK Government in a quite different light from the one in which it presented itself at the Global Conference on Media Freedom:

> "We have significant concerns over the scope of harms included as well as the model being proposed, and the risks that they would pose to the rights to freedom of expression and privacy," wrote Global Partners Digital in a statement that was echoed by many other groups. "We believe that the proposals, if taken forward in their current state, would likely put the UK in breach of its obligations under both international human rights law and the European Convention on Human Rights (ECHR), as incorporated into domestic law through the Human Rights Act 1998 (HRA 1998)."

How has the UK government's split policy personality on freedom of expression come about? Can anything be done to resolve it?

Part 2 of this paper explores how the manner in which the UK Government has framed the challenge of online disinformation risks damaging the very ideals it espouses in its domestic and foreign policy, and then sets out ways the UK Government can reframe its approach in order to play the global role it has set for itself as defender of freedom of expression.

The UK White Paper on Online Harms shows that the debate on regulating online "disinformation" has reached a critical fork in the road. Along one path lies an opportunity to strengthen freedom of speech, human rights and deliberative democracy for the 21st century; down another lies the risk of fundamentally misunderstanding the nature of the internet, damaging freedom of speech, and imposing a political logic and language that favors authoritarian regimes. While the White Paper has some encouraging ideas that hint at the possibility of taking the former path, its overall language and framing show that the UK Government risks stumbling, perhaps unthinkingly, down the second.

The Censorial Cycle

The UK White Paper is right to demand that tech companies take more responsibility for illegal material and behavior on their platforms. From Facebook-fueled ethnic cleansing in Myanmar to the continued circulation of child pornography online, tech companies need to comply with national and international law and with international human rights legislation especially. The White Paper is also wise in not making companies liable for every piece of content on their platforms, which is almost impossible to enforce technically, but to demand they put in place the systems to mitigate the dissemination of illegal content.

However the White Paper is on much thinner intellectual ice when it comes to the more delicate question of what it calls "harmful but legal" content, which it intends to lump together under the same "duty of care" as outright illegal material, and which includes everything from online bullying to the subject of this paper, disinformation.

"Disinformation" is not a legal concept, and clamping down on it unavoidably runs counter to international legislation on freedom of expression. As the media law scholar Damian Tambini argues,

> The central legal and constitutional problem here is that establishing new standards in a code of conduct, and introducing sanctions and fines for "harms with a less clear definition" and that are also legal, does not pass the European Convention on Human Rights free speech test according to which restrictions have to be prescribed by law, and necessary, for a legitimate aim. The requirement that restrictions should be "prescribed by law" is a safeguard against a slippery slope to censorship. Constraints on speech should not be imposed on the basis of opaque agreements between platforms and politicians – a scenario arguably left open by the White Paper – they should be subject to the constraint of parliamentary debate.[16]

As many free speech groups have noted, the extensive list of punishments companies risk being subjected to if they do not comply with the duty of care (which include fines, blocking their business from operating, even imprisonment) risks a "chilling effect" on free speech, where tech companies would rather take material down than face the risk of such punishment. Though the White Paper does make some noises about protecting freedom of expression, in practice its proposals show scant respect for it. For example, the White Paper notes the need for a "super complaints" procedure for users to demand action from technology companies when they do not abide by their Duty of Care. One might expect such a "Super Complaints Procedure" to primarily protect people whose content has been taken down to seek redress. Instead the White Paper stresses that the super complaints procedure should be used for the opposite, as a way for individuals to demand companies take material down.

But there is more at stake here than just complaints procedures around content moderation. As currently formulated the White Paper's logic and language are skewed toward suppression rather than defense of freedom of expression. The White Paper invokes an idea of speech as somehow inherently dangerous, something that people need to be protected from. This is a logic and language that will suit those authoritarian regimes, such as Russia, that aim to frame the debate around internet regulation in such a way as to legitimize censorship, to create what Russian and Chinese advocates of censorship call a "sovereign internet" where they can control content and slow down the free flow of information across borders.[17] As David Kaye, the UN Rapporteur on Freedom of Speech and a professor at the University of California, Irvine, says: "A 'rhetoric

of danger' is exactly the kind of rhetoric adopted in authoritarian environments to restrict legitimate debate, and we in the democratic world risk giving cover to that."

There is a distinct irony at work here: one of the motivations for introducing regulation in the UK has been the covert online campaigns waged by the Kremlin to influence democratic processes in the U.S. and Europe. Now the UK's response risks imposing exactly the sort of ideas for governing the internet the Kremlin is promoting.

One could argue that authoritarian regimes will enact censorship irrespective of how democracies regulate the internet. This may be the case, but the mission of democracies should be to propose an alternative regulatory vision, one that empowers and inspires in line with human rights and strengthens democratic ideals worldwide. This is especially true of the UK, which sees itself as a global leader in setting standards for media freedom and information policy. When creating regulatory proposals policy makers should therefore always be asking themselves what is the difference between an authoritarian internet and a democratic one, how do regulatory proposals strengthen the latter or at least avoid damaging it. By taking on the language and logic of authoritarian regimes, the UK Government risks reinforcing a censorial cycle: the more covert online influence campaigns authoritarian regimes such as Russia launch in democracies, the more democracies adopt a frame and policy logic these regimes favour.

We already see this cycle being perpetuated in the raft of policy proposals across the world to combat "fake news." A German law to take down "illegal" content – including blasphemy – has been quoted by Russia and Singapore as they put forward their own punitive legislation. The Singapore law is committed to fighting "fake news" and "disinformation" – but as defined by the government.[18] Journalists fear it will be used as an excuse to attack them.

In *Don't Think of an Elephant!,* the cognitive linguist George Lakoff defines winning and losing in politics as being about framing issues in a way conducive to your aims. Defining the argument means winning it. If you tell someone not to think of an elephant, they will end up thinking of an elephant. "When we negate a frame, we evoke the frame…when you are arguing against the other side, do not use their language. Their language picks out a frame – and it won't be the frame you want."[19]

But even as one rejects the "rhetoric of danger" and the negative framing the White Paper shares with authoritarian regimes, one needs to also recognize how online disinformation is qualitatively different than older forms. While producing erroneous content is not new, technology now makes it possible to disseminate content at unheard-of rates, targeted at specific audiences. As the law professor Tim Wu argues:

> The most important change in the expressive environment can be boiled down to one idea: it is no longer speech itself that is scarce, but the attention of listeners. Emerging threats to public discourse take advantage of this change… emerging techniques of speech control depend on (1) a range of new punishments, like unleashing "troll armies" to abuse the press and other critics, and (2) "flooding" tactics (sometimes called "reverse censorship") that distort or drown out disfavored speech through the creation and dissemination of fake news, the payment of fake commentators, and the deployment of propaganda robots."[20]

Speech itself, argues Wu, is being used as a "censorial weapon."

So how can one simultaneously respond to the specific challenge of digital era disinformation, while strengthening democratic ideals and freedom of expression?

Frame 1: Don't Mention It

To achieve the minimal aim of not damaging freedom of expression with its regulatory proposals, the UK Government can simply avoid regulating legal (if untrue) speech as a category in and of itself, and instead expand on existing legislation sector by sector and environment by environment to deal with disinformation in specific contexts. This is a sort of framing by omission, where the "disinformation" question is not dealt with as a separate category, but becomes a subset of other legislation.

An obvious place to start is electoral advertising. Current regulations around transparency and accuracy of political advertising and election integrity focus on traditional print and broadcast media. These must be updated to address the new reality of online political microtargeting, where adverts are created in the millions with different messages aimed at niche audiences. There needs to be a legal requirement to create an easily searchable repository for all election-related ads in real time that clearly shows who has paid for them, to whom they are targeted, and which of a person's data is used to target them. Moreover, as the Coalition for Reform of Political Advertising and Incorporated Society of British Advertisers propose, all factual claims in the political ads should be pre-cleared and the ads should be watermarked to show their origins.[21] As political parties in the UK pulled out of regulation from the Advertising Standards Authority, essentially, in the neat phrase of Full Fact, "political parties have chosen to hold themselves to lower standards than washing powder sellers," this regulatory function will have to be placed under a body such as auspices of, for example, the Electoral Commission.

The challenge, of course, is whether the current definition of "electoral ads" is sufficient. Electoral law currently defines electoral material as "material which can reasonably be regarded as intended to promote or procure electoral success at any relevant election." Just a glance at the current environment in the UK, where non-transparent online ads are being used to influence the Brexit debate as we speak, shows how online ads are being used all the time to shape political outcomes. All ads by political parties and government agencies should show full information on content, targeting reach and spend, which must all be provided in real time. This still, however, would not cover the full spectrum of political ads, such as issue-based ads by civic groups, proxies and allies. Indeed, given there is no settled scope of what a "political" ad is, ultimately all paid-for content should have this level of transparency attached.

Elsewhere existing legislation on public health could be drawn on to ensure that people are informed of health risks propagated by inaccurate online disinformation about, for example, vaccines. As to foreign interference campaigns, where not covered by regulation around election integrity and political advertising, the most egregious cases could be addressed under national security policy. Importantly, national security policy not only comprises legislative and regulatory measures, but may also engage diplomatic channels for the resolution of foreign interference. The pushback against covert foreign campaigns might not be in the information space at all, but in asymmetric responses such as economic sanctions against hostile states.

Election, privacy, national security and public health are not an exhaustive list of sectors that need to be updated for the digital age, but they are obvious examples of how a sectoral approach would deal with specific aspects of "disinformation" without having to impose blanket bans on types of speech. As discussed, content

moderation of "disinformation" will invariably bring a collision with freedom of speech. It is also largely impractical, encouraging a "whack-a-mole" approach. Most importantly it fails to understand that information operation campaigns, such as the infamous Russian Internet Research Agency campaign in the U.S., can use neutral or even accurate content in their activity. Much of the Russian covert U.S. campaign, for example, simply supported various causes and politicians, without giving any specific accurate or inaccurate facts. Rather than deceptive content, these campaigns are marked by what Facebook calls "coordinated inauthentic behavior" or what one could term "viral deception": where the actors disguise both their true identity in order to deceive people, and where material is promoted in an inauthentic way to make it look more popular than it is.

Frame 2: From Content to Behavior

If we reframe "disinformation" as pertaining less to content and more to behaviour – here, referring to (artificial) technical means to boost the dissemination of certain content – we get away from the problem of regulating speech and onto the systemic use of technology to deceive people, from regulating statements to focusing on the use of bots, cyborgs and trolls that purposefully disguise their identity to confuse audiences; cyber-militias whose activity seems organic but who are actually part of planned campaigns full of deceptive accounts; the plethora of "news" websites free of journalistic standards that look independent but are covertly run from one source, all pushing the same agenda. The issue here is not anonymity, which is sometimes necessary to guarantee safety, but the right of people to understand how the information environment around them is being shaped. Shouldn't one have the right to know if what looks organic is actually orchestrated? How the reality one is interacting with is engineered? Shouldn't bots, to give a small example, always be clearly marked as bots?

A first legislative step in this direction has been taken in California, where bots are now forced to reveal their "artificial identity" when they are used for commercial or electoral purposes. "It's literally taking these high-end technological concepts and bringing them home to basic common-law principles," Robert Hertzberg, a California state senator who is the author of the bot-disclosure law, told the New Yorker. "You can't defraud people. You can't lie. You can't cheat them economically. You can't cheat 'em in elections."[22]

Could such transparency around online behavior be taken further? Can we imagine an online life where any person would be able to understand how the information meteorology around them is being shaped; why computer programs show you one piece of content and not another; why any ad, article, message or image is being targeted specifically at you; which of your own data has been used to try and influence you and why; whether a piece of content is genuinely popular or just amplified. Ideally such information should be instantly available in real time, so that, for example, one could click on or hover over a piece of online content and be able to immediately access its provenance. Ultimately different technological solutions will need to be found for different platforms. What matters is framing the issue in a way that demands more information, not censorship, which empowers the user. Maybe then we would become less like creatures acted upon by mysterious powers we cannot see, made to fear and tremble for reasons we cannot fathom, and instead would be able to engage with the information forces around us as equals.

Such a reframing of the "disinformation" debate to focus on uncovering deceptive behaviour and empowering a person online to understand the information environment around them takes us away from the logic and

language that authoritarian regimes promote. We are back in a framing that increases people's rights and freedoms, rather than constricting them. As David Kaye, the UN rapporteur on Freedom of Expression, told me:

> Another way to conceptualize the impact and purpose of viral deception – assuming we can define it sufficiently narrowly – is as a tool to interfere with the individual's right to information. Coordinated amplification has a strategic aim: make it harder for individuals to assess the veracity of information. It harms public debate, but it also interferes with the individual's (per A19 of the ICCPR/UDHR) "right to seek, receive & impart information and ideas of all kinds." Conceived this way, it seems to me that solutions could be aimed at enhancing individual access to information rather than merely protecting against public harm.[23]

Internet Transparency as Unifying Principle for Democracies

Focusing on disinformation as pertaining to "inauthentic" behaviour in disseminating content helps open a broader discussion about the transparency of the internet, an approach that freedom-of-expression advocacy organisations are more comfortable with than regulating content itself. As Article 19 write in their response to the White Paper:

> …the regulator should not be involved in the determination of the legality of content, but instead focus on transparency obligations and reviewing internal company processes on content moderation…
>
> Government should focus on greater transparency and accountability mechanisms in the application of companies' terms of service/community standards … digital companies should explain to the public how their algorithms are used to present, rank, promote or demote content. Content that is promoted should be clearly marked as such, whether the content is promoted by the company or by a third-party for remuneration … they should publish information about the methods and internal processes for the elaboration of community rules.[24]

This focus on "transparency" has at least three advantages.

First, it puts the focus on empowering people online, augmenting their right to receive information, rather than constricting freedom of expression.

Second, greater transparency will help balance the information field to give public interest news organisations, such as fact-checking agencies, a fighting chance.

Though there have been some tenuous efforts by social media companies to work with fact-checkers, the relationship is deeply unequal and the lack of transparency makes it hard for fact-checkers to operate with efficiency. In interviews I have conducted with fact-checkers, for instance, many complain they have little knowledge about which pieces of disinformation they should focus on, as they cannot see which pieces of disinformation are being aggressively amplified by political actors. Instead they can spend time and effort on debunking ineffectual lies. If there were enough transparency to show which pieces of disinformation are being pushed through targeted, coordinated, inauthentic campaigns, then this would signal to the fact-checkers which content to focus on.

Third, and perhaps most important in the context of this text, internet transparency can become the consensus position that democracies can agree on as a unifying principle for both regulating the internet and promoting democratic values, one that stands in robust contrast to the mix of censorship and non-transparency that define authoritarian approaches. This is a way for the UK to be both a global leader in promoting media freedom and promoting a way to regulate the internet that protects freedom of expression. When the next Global Conference on Media Freedom rolls around, the UK will be able to trumpet how its vision for the internet is in harmony with its noble support to protect journalism.

Notes

[1] Peter Pomerantsev is a Senior Fellow at the Institute of Global Affairs at the London School of Economics and Political Science, where he is Director of the Arena Initiative. An author and TV producer, he specializes in propaganda and media development, and has testified on the challenges of information war to the U.S. House Foreign Affairs Committee, U.S. Senate Foreign Relations Committee and the UK Parliament Defense Select Committee. He is the author of *This Is Not Propaganda: Adventures in the War Against Reality*, published in August 2019 by PublicAffairs, and *Nothing Is True and Everything Is Possible* (2016), which won the Royal Society of Literature Ondaatje Prize.

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

[3] https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm

[4] Ofcom, September 2018, "Addressing Harmful Online Content"

[5] Woods, Perrin. Carnegie UK, January 2019, "Internet Harm Reduction"

[6] Full Fact, 2018, "Tackling Misinformation in an Open Society"

[7] To note: UK Music has already called upon the British Government to expand the scope of the paper to protect the culture and creative industries.

[8] https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm

[9] https://www.indexoncensorship.org/2019/04/uk-government-online-harms-white-paper-shows-disregard-freedom-expression/

[10] https://www.article19.org/resources/uk-article-19-response-to-leaked-reports-on-online-harms-white-paper/

[11] https://www.spectator.co.uk/2019/04/iplod-sajid-javids-new-internet-rules-will-have-a-chilling-effect-on-free-speech/

[12] https://www.cyberleagle.com/

[13] https://www.carnegieuktrust.org.uk/publications/internet-harm-reduction/

[14] https://www.gov.uk/government/topical-events/global-conference-for-media-freedom-london-2019/about

[15] https://www.gov.uk/government/speeches/addressing-threats-to-media-freedom-amal-clooneys-speech

[16] Reducing Online Harms through a Differentiated Duty of Care: A Response to the Online Harms White Paper | FLJS https://www.fljs.org/content/reducing-online-harms-through-differentiated-duty-care-response-online-harms-white-paper

[17] https://www.hrw.org/news/2019/04/24/joint-statement-russias-sovereign-internet-bill

[18] https://www.nybooks.com/daily/2019/07/19/singapore-laboratory-of-digital-censorship/

[19] Lakoff, George. *Don't Think of an Elephant!* Chelsea Green, 2014

[20] https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete

[21] https://coinform.eu/wp-content/uploads/2019/02/Full_fact_tackling_misinformation_in_an_open_society.pdf

[22] https://www.newyorker.com/tech/annals-of-technology/will-californias-new-bot-law-strengthen-democracy

[23] https://www.the-american-interest.com/2019/06/10/how-not-to-regulate-the-internet/

[24] https://www.article19.org/wp-content/uploads/2019/07/White-Paper-Online-Harms-A19-response-1-July-19-FINAL.pdf