INTERNET POLICY REVIEW
Journal on internet regulation

Volume 10 **Issue** 2

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

The promise of financial services regulatory theory to address disinformation in content recommender systems

Owen Bennett *Independent*

DOI: https://doi.org/10.14763/2021.2.1558

Published: 11 May 2021
Received: 13 November 2020 **Accepted:** 25 February 2021

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Bennett, O. (2021). The promise of financial services regulatory theory to address disinformation in content recommender systems. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1558

Keywords: Disinformation, Regulatory theory, Content recommender systems, Digital Services Act

Abstract: This article argues that the European regulatory approach to disinformation online is stymied by inappropriate regulatory theories. On that basis, this article seeks to advance an alternative theoretical approach, inspired by the contemporary European paradigm of financial services regulation. It outlines how the key theories underpinning financial services regulation could engender policy solutions that are both more rights-protective and more responsive to the role played by content recommender systems in compounding the policy problem of disinformation online. It assesses the extent to which these alternative regulatory theories manifest in the draft EU Digital Services Act.

Section 1. Introduction

Across Europe, policymakers and the public are increasingly demanding the regulation of so-called 'harmful-but-legal' online content (European Commission, 2018). Disinformation —information that is false and deliberately created to harm a person, social group, organisation or country—is a quintessential example of this 'harmful-but-legal' phenomenon (Wardle & Derakhshan, 2017, p. 20). [1] The motivation for regulatory interventions to address disinformation online is well-founded. There is an ever-increasing body of empirical evidence that links it to 'real' harms; harms that cut across individual welfare, broader social interests, and democratic stability (Shmargad & Klar, 2020; Vaccari & Chadwick, 2020). The ongoing COVID-19 pandemic bears witness to this unsettling reality, with online disinformation being linked to a variety of untenable outcomes—from attacks on telecommunications infrastructure, to the consumption of dangerous 'miracle cures', to increased xenophobia.

Yet despite the broad desire for intervention, policymakers in the EU and UK have thus far struggled to develop effective regulatory responses. This results from the fact that the long-standing paradigm within which content regulation in Europe manifests, and thus the starting-point for discussions concerning the regulation of disinformation online, is underpinned by regulatory theories that are unsuitable for the problem at hand. As will be explained in detail in the following sections there are two principle and related reasons for this. First, the regulatory theories underpinning our contemporary approach take their objective to be suppression of *content* that is considered objectionable, a reality which ignores the wholly *context-dependent* nature of the harm in disinformation. Second, much of the causal factors that give rise to the *policy problem* of disinformation online are informed by the *business practices* of certain contemporary online service providers, most notably their provision of open content recommender systems. Yet in spite of this, the theories that underpin the contemporary approach to content regulation largely ignore business practices by firms in the market as potential sites for regulatory intervention.

In that context, the aim of this article is to open the door to a potentially more effective approach for addressing disinformation online, by borrowing from contemporary European financial services regulation. [2] The intention of this article is not,

1. This should not detract from the fact that some content that falls within the broad definition of disinformation may indeed be illegal under national law, for instance disinformation that meets the standard of proscribed hate speech or which incites violence.

2. Where utilised in this article, 'European' refers to the legal frameworks of both the European Union

however, to set out a fully-fledged alternative regulatory framework. Rather, it explores whether and to what extent the *theories* that underpin financial services regulation could be borrowed by policymakers who seek to develop a more appropriate regulatory response to disinformation, and indeed how they could serve as the crucial *underpinning* of novel regulatory interventions. To frame that endeavour, section 2 explains precisely *why* the present regulatory paradigm and its underlying theories are unsustainable, with a specific focus on their treatment of freedom of expression and content recommender systems. Section 3 isolates the key theories that underpin financial services regulation and outlines how they might be utilised to underpin a novel regulatory approach to disinformation online. Section 4 then moves to illustrate the improvements that such an approach would have over the status quo, while section 5 engages with the potential shortcomings. Section 6 concludes by plotting a course forward for this area, in context of the recently-proposed legislative proposal for an EU Digital Services Act ('DSA').

This article sits within the emergent body of scholarly and policy work that seeks to identify next-generation approaches to the governance of online content and the regulation of large content-sharing platforms. Many of these proposals highlight the need for an appreciation of, and regulatory attuning to, power dynamics in the platform ecosystem (Helberger, 2020; Graef & van Berlo, 2020; Gillespie et al., 2020); others identify firms' business practices and commercial logics as motivations for, and necessary sites of, regulatory intervention (Cobbe & Singh, 2019; Woods & Perrin, 2019; Gary & Soltani, 2019); and others again provide paths forward for scrutinising and evaluating said practices and power dynamics (Wagner et al., 2021; Leerssen, 2020). As such, these efforts anticipate the kind of regulatory *solutions* that are required to better address the problem of disinformation and other online harms facing the internet ecosystem today. The contribution that this article seeks to make is to provide the necessary suite of regulatory theories that can underpin those promising solutions.

Ultimately, my contention is that a regulatory approach that grounds itself in the theories of financial services regulation would help us to better moderate the business practices that can make disinformation harmful, while engendering less interference with fundamental rights. However, this focus on addressing disinformation online through platform regulation is not premised on reductive technological determinism. We will not 'solve' disinformation by regulating online content. Problem definitions and policy solutions must not ignore the political, sociological, and economic contexts and structures within which disinformation online emerges.

and the United Kingdom.

Nor should they ignore the role of entities that develop, transmit, and amplify disinformation for malign ends. Simply put, this is a multifaceted problem that necessitates multiple vectors of intervention both online and off. Some of the most important interventions—particularly those in the domain of media literacy—require action and investment now but whose 'pay-off' may not be observable for several years. That I have chosen to focus my attention on one component is not intended to dismiss the need for a holistic interdisciplinary approach.

Section 2. Disinformation online and the regulatory context

2.1 A focus on the wrong target

As noted above, the contemporary regulatory paradigm within which our approach to disinformation online is situated systematically engenders unjustified interferences with individuals' fundamental rights, in particular freedom of expression. This results from the fact that it takes its objective to be the regulation of content *as such*, and more precisely, from its strategic reliance on what is known in regulatory theory as performance and technology-based approaches to regulatory intervention (Coglianese & Lazar, 2003). In practice, these theoretical underpinnings typically manifest in law as output targets (e.g. the removal by firms of all notified content within 24 hours) and specific technological mandates (e.g. the deployment by firms of automated content filters). Crucially, performance and technology-based interventions aim at achieving certain outcomes with respect *to content that is objectionable*. Put another way, the belief is that these are 'content' problems, and as such, warrant 'content' solutions (Francois, 2019, p. 2).

One key reason why this regulatory strategy gives rise to systematic and intolerable rights interferences when applied to disinformation online is because the 'harm' of disinformation is not to be found in some objective and essential feature of the content. On the contrary, the fact of whether disinformation is harmful by any given metric of 'harm' is wholly *contextual*. It depends on the intersection of various factors related to the content, the consumer, and the broader social, political, and cultural environments within which that content is consumed (Wardle & Derakhshan, 2017). As such, parsing 'harmful' disinformation from trivial falsehoods, satire, and unintended inaccuracies is no easy task, and certainly not one that can be made with reference to the content alone (Francois, 2019). Policymakers are thus left in a bind. The regulatory paradigm within which they operate assumes harm to be located predominantly *within content*, and therefore the toolbox at their disposal includes almost exclusively instruments that address *content as*

such. 'Compliance success' under this 'content-centric' approach means *more* take-downs and *more* filtering. Yet given that identifying disinformation requires careful assessment and its 'harm' depends on a variety of contextual factors, an approach which optimises for speedy content suppression at scale is naturally going to lead to systematic over-removal and suppression of legitimate expression, often with little possibility for affected individuals to plead their case or seek redress (Keller, 2019; Engstrom & Feamster, 2017). To give individuals' fundamental rights the protection they warrant, we need a different approach.

2.2 Blind to the business practices

Yet not merely rights-interfering, the contemporary paradigm is also *ineffective* at addressing disinformation online. This is a consequence of its silence with respect to the influential role played by the business practices of certain types of online service providers (henceforth, 'online service providers' or 'OSPs') in exacerbating the policy problem. Indeed, certain types of OSPs control many of the aforenoted contextual factors that determine whether a piece of disinformation is absorbed by its consumer as either a trivial falsehood worthy of ridicule or as a serious source of 'real-world' harm (Cobbe & Singh, 2019). This occurs most notably where they operate content recommender systems.

Following Cobbe & Singh, I define content recommender systems as product features that *algorithmically* rank and present content to particular users, according to some determination made by the OSP of the relevance, interest, and importance of the content for that particular user (Cobbe & Singh, 2019, p. 3). Content recommender systems come in many forms, but of particular interest for our purposes is the 'open' model, such as Facebook's News Feed and YouTube's 'Up Next' feature, whereby *unvetted* third-party content is selected for promotion to a new audience (henceforth, 'open content recommender systems' or 'OCR systems'). [3] In determining what third-party content to surface for a user, the system's algorithm typically draws upon user inputs (e.g. what that individual has already consumed); group predictors (e.g. what similar types of users have consumed); and a variety of other contextual personal data points (e.g. profiling data from data brokers) (Ibid, 2019). As such, OSPs that operate OCR systems (hereafter, 'the OCR sector') largely define *what* content is seen, *whom* it is seen by, and *how* it is presented, while simultaneously being subject to no formal editorial control obligations vis-à-vis said content. [4]

3. This is of course notwithstanding the fact that some OSPs do engage in pre-screening of content in certain instances, particularly with respect to copyright infringement (e.g. YouTube's Content ID) and child sexual abuse material (e.g. Microsoft's PhotoDNA).

The problem is that the commercial incentives that determine the design and operation of these systems may inadvertently exacerbate the spread and impact of disinformation (Marechal & Roberts Biddle, 2020; Singh, 2019; Gary & Soltani, 2019). In short, the business model of the major OSPs operating OCR systems is targeted advertising; the OSPs' revenue function depends on consumers spending time on the platform and consuming advertising content. OCR systems help maximise this revenue function by presenting users with the kind of personalised content that keeps them engaged with the platform, and hence addressable with advertising content (Solsman, 2018). Yet much of the content that typically engages users is shocking and misleading, and hence OCR systems that are designed to maximise user engagement can inadvertently compound the problem of disinformation and other 'harmful-but-legal' content (Gary & Soltani, 2019, §1). For instance, internal research from Facebook, undertaken in 2016 and only recently brought into the public domain by investigatory reporting, found that '64% of all extremist group joins are due to our recommendation tools' (Horwitz & Seetharaman, 2020, §2). As Balkin (2018, p. 3) observes, 'the same business model that allows companies to maximize advertising revenues also makes them conduits and amplifiers for propaganda, conspiracy theories, and fake news'.

Unfortunately, complex OCR systems and their contribution to the problem of disinformation were unforeseen at the time that the present European regulatory paradigm was coming into being. It makes *content* the object of regulatory intervention, yet is largely silent on the systems and processes by which that content is served and consumed. In the case of disinformation this limitation poses acute challenges. Given that the harm of disinformation depends wholly on contextual factors—many of which are controlled by OSPs that operate OCR systems—it is untenable that the regulatory approach should remain passive with respect to the design and operation of these systems. On the basis of these shortcomings we need to change course in our regulatory approach to disinformation online. Yet we find few obvious alternative solutions within the domain of online content regulation. I therefore propose that we look further afield, to a sector with a longer tradition of intensive regulation, namely, the financial services sector.

4. Admittedly this will soon change to an extent as, under the revised EU Audiovisual Media Services Directive, 'video-sharing providers' will for the first time be subject to certain principle-based editorial obligation measures. It should also be stressed that the fact that OCR systems remain outside the scope of traditional editorial obligations that apply to broadcasters is not a policy weakness *per se*. As Leerssen (2020) points out, unlike in the broadcasting realm users are not wholly passive with respect to open content recommender systems—in almost all cases the content served to the user is *somewhat* informed by those users' explicit and implicit preference signaling.

Section 3. Imitation is the greatest form of flattery: learning from the financial services sector

3.1 Methodological motivations and justifications

The financial services sector is a suitable candidate of comparison for a number of reasons. As Black (2012) notes, it has been the testing ground for many of the leading *new governance* theories of regulation over the last 30 years. As such, by looking to the world of financial services we can grasp a landscape picture of the various possible regulatory theories at our disposal. The motivation for the *substantive* comparison arises because the factors that shape regulatory dynamics in the financial services sector appear similar to the OCR sector in three important respects: first, the structure of the market; second, the nature of firms' incentives; and, third, the consequences that arise when things go wrong.

In the first case, the financial services sector is characterised by various types of actors (e.g. from retail to investment banking), operating at different scales (from multinationals to credit unions), and involved in numerous lines of business (from deposit holding to mortgages to wealth management). As such, there is a high degree of market heterogeneity, with regulation evolving in response to this. Considerable heterogeneity can also be observed in the OCR sector, where firm size varies and where OCR systems take on distinct roles within broader product bundles. For instance, Facebook's News Feed OCR system is a core feature of the user experience architecture, serving as a pathway to 'ancillary' services such as interest groups as well as public pages. In contrast, YouTube's OCR system takes multiple distinct forms, including AutoPlay and homepage recommendations. Crucially, in both instances the firm's OCR system is bundled within largely distinct online services that could arguably be said to operate within different product markets. The revenue base for firms in the broader OCR market is likewise varied, even amongst the largest actors. For instance, Twitter's 2019 revenue amounted to US$ 3.46 billion, a figure dwarfed by Facebook's US$ 70.7 billion (Macrotrends, 2021).

In the second case, a guiding assumption of financial services regulation is that firms' commercial incentives are naturally misaligned with the public interest, and so regulatory intervention is essential to ensure that firms will internalise costs that would otherwise be externalised (House of Commons, 2009). As we saw in the section prior, an increasing body of research suggests that a similar dynamic of misaligned incentives is at play in the market for OCR systems. Therein, the commercial incentives underpinning the design and operation of OCR systems can inadvertently exacerbate the spread and impact of disinformation. Although we do

not yet possess a comparable degree of insight as in the financial services sector, with every passing day new evidence emerges that points to a pronounced structural misalignment between the OCR sector's incentives and the broader public interest (Hao, 2021; Horwitz & Seetharaman, 2020; Bergen, 2019).

In the third and final case, the comprehensiveness and intensity of the financial services regulatory regime is a response to the sheer degree of harm that may arise when financial services firms act in a manner that is contrary to the public interest (e.g. loss of savings by individuals and businesses; exploitation of vulnerable consumers, etc.). While the types of harm that may arise from disinformation in OCR systems (e.g., individuals ignoring crucial public health messaging during a pandemic; groups of voters succumbing to voter suppression efforts during an election period) are obviously different in nature, their impact in terms of negative public interest outcomes can arguably be similar. Indeed, it is not without reason that the draft DSA identifies OCR systems as a vector for 'systemic risks' and hence worthy of heightened oversight (COM/2020/825 final, rec. 54).

Equipped with this comparative framing, we may now turn to the question of what precisely are the key regulatory theories underpinning financial services in the EU and UK and how they might be applied to our domain. Broadly speaking, European financial services regulation includes three theoretical features that I believe could underpin a more effective policy response to disinformation online: first, a comprehensive focus on risk management; second, a dependence on principle-based rules; and third, a reliance on regulated firms to achieve desired policy outcomes (Black, 2012). While there are differences across jurisdictions and there is of course more to the sector's regulatory theory than these three elements, these are selected for consideration on the basis of both their foundational role within financial services regulation and their *prima facie* promise for our purposes. In what follows, I will briefly outline their meaning before articulating their envisaged application.

3.2 Thinking in terms of risk

'Risk' in the financial services regulatory theory should be understood broadly, as the concept serves to underpin numerous modalities of policy therein. Most notably, risk provides the basis for regulatory *legitimacy*. Financial institutions are understood as posing systemic risks of various kinds, and that if left to their own devices firms are incapable of identifying, managing, and mitigating those risks. This understanding motivates and legitimises a body of law—known as 'prudential regulation'—that intervenes with, and deeply scrutinises, everything from firms' commercial practices, to their risk-mitigation strategies, and even the composition of

their senior leadership. For instance the EU Prudential Requirements directive grants regulatory authorities the power to limit or prevent business practices which pose 'excessive risk to the soundness of an institution' (Directive 2013/36/EU, art. 92.2 (a)), and to take steps to ensure banks' remuneration policies 'promote sound and effective risk management and do not encourage [excessive] risk-taking' (Ibid, art. 104.1(e)). In addition, risk often manifests as the *metric* by which compliance measures are determined. For instance, the body of law concerning anti-money laundering and terrorist financing (AML-TF) is grounded in the belief that it is impossible to ever fully prevent a firm's services from being exploited to launder money or finance terrorist activities. Consequently, the counter-measures that a regulated entity should take to address these unlawful practices—such as know-your-customer due diligence and transaction analysis—are to be *commensurate* with the *risk* of the exploitation occurring. Entities like the Financial Action Task Force issue regular authoritative compliance guidance for regulators and firms on how to assess, evaluate, and manage AML-TF risks in practical settings.

To utilise the risk-based approach as an underpinning of our policy response to disinformation online, we would of course first need a clear *conceptual understanding* of risk as it pertains to our domain. Specifically, policymakers would be required to establish a methodology for determining various individual and public interest disinformation-related risks as well as a hypothesis explaining how these risks can manifest in both online content and on the basis of the commercial practices of firms. An effective risk schema should provide the means by which—in different contexts—we could assess whether and to what extent a given piece of disinformation is likely to cause harm to the public interest, and likewise assess the risk profile of specific commercial practices and product features.

Policy interventions on the basis of this approach would aim at the identification, management, and mitigation of these disinformation-related risks. Crucially, under the risk-based approach compliance efforts would be *commensurate* with the degree of risk posed to the public interest—this focus on 'commensurability' means that the fact of *hosting* disinformation would not *in itself* be indicative of noncompliance with the law. Moreover, the adoption of a risk-based approach would require a general shift in the locus of policy intervention *away* from content removal. Today, intervention often aims at firms' *outputs*, meaning action occurs once the risk is *materialising*. Yet, many of the risks associated with disinformation are intimately related to firms' business practices and how they engage with third-party content on their services. Consequently, a true risk-based approach to addressing disinformation would manifest in regulatory interventions *earlier* in the product cy-

cle. This means focusing particularly on OCR systems, given their significant influence in shaping the spread and impact of disinformation. For instance, it could manifest as regulatory obligations regarding algorithmic auditing and inspection, that aim at continuous quality assurance of OCR systems and early warning of any operational flaws. In addition, it could manifest as measures that place restrictions on the micro-targeting of content to specific types of uses, given that much of the public interest risk in disinformation is determined by *who* sees the content and under *what* circumstances (Haines, 2019).

Notably, the focus on 'risk' shines through in the draft DSA in at least two distinct modalities. First, we have seen how risk serves as the basis for regulatory legitimacy in the financial services sector. The DSA adopts a similar approach, delineating a category of online service providers—so-called 'Very Large Online Platforms'—whose size, influence, and risk justifies asymmetric regulatory obligations (COM/2020/825 final, art. 25). In addition, risk manifests as a metric of compliance in a manner similar to AML-TF regulation, with firms obliged to assess "the *systemic risks* stemming from the functioning and use of their service, as well as by potential misuses by the recipients of the service, and take *appropriate mitigating measures*" (Ibid, rec. 56, my emphasis). Notably, while the provisions of the DSA that establish direct obligations and responsibilities with respect to content—e.g. the provisions on notice & action—limit their focus to that which is *illegal*, the risk assessment and mitigation measures enshrined in the law's articles 26 and 27 *do* implicate disinformation explicitly and implicitly. Indeed, these provisions draw out 'coordinated inauthentic behaviour' as a risk to be mitigated (Ibid, art. 26. 1 (c)) and likewise earmark OCR systems as a site where risk mitigation measures should be located (Ibid, art. 27.1 (a)).

Yet we have discussed that for risk to serve as an effective basis for regulatory intervention (in the case of disinformation or any other 'harmful-but-legal' content), policymakers must first establish a rigorous risk schema. In that regard, the DSA is somewhat underwhelming, in that its guidance on what risks to assess and how they ought to be mitigated is vague. In the case of risk assessment, companies are expected to translate interferences with fundamental rights into the language of risk management, with little delineation of how those risks might be understood or quantified in practical terms. As we will discuss in section 5, this shortcoming brings a myriad of theoretical and practical challenges. In addition, risk is typically understood across domains as a function of the probability and severity of a certain (inverse) outcome (ISO, 2018 §3). Yet the legal architecture of the DSA's risk-based approach appears overly weighted toward addressing the *severity* of risks

rather than their probability, notably through its focus on addressing systemic risks related to 'dissemination' of illegal content and the 'intentional manipulation' of their service. While of course it is important to assess and address the severity of risks that materialise (e.g. through developing partnerships with trusted flaggers; deploying content recognition software to assist human review), risk assessment in this domain will only be meaningful if it gives equal weight to the *probability* of a certain risk materialising (e.g. assessing whether OCR system design may inadvertently privilege the virality of disinformation).

3.3 Principles over prescription

A second key feature of regulatory theory in the European financial services domain, most associated with the UK, is the emphasis on principles-based rule-structures. The rules governing the conduct of regulated entities are often 'general, qualitative, purposive, and behavioural' (Black, 2008, p. 13). A quintessential example of this approach is found in the UK Prudential Regulatory Authority's 'Fundamental rules', where regulatory obligations are expressed in such terms as 'a firm must organise and control its affairs responsibly and effectively' (PRA, 2014, p. 5). In the EU, the 2013 Prudential Requirements directive exhibits similar principles-based requirements with respect to the regulation of business conduct, demanding that regulated firms to have in place 'robust', 'sound', and 'effective' policies and governance arrangements (2013/36/EU, art. 73). Transitioning from the macro to the micro—that is, the process whereby the principles are translated into specific compliance approaches for individual firms—is usually done through iterative dialogues between firms and regulators as well as through guidelines and secondary legislation.

While admittedly the regulating-by-principles approach can already be found in some foundational elements of the EU and UK content regulation legislative *acquis*, the paradigm has become increasingly defined by 'prescriptive' and 'bright-line' rule forms in the last two decades, whereby rules often take the form of highly-complex descriptive obligations or simple rigid directives respectively. [5] To understand how principles-based rule-structures could be more formally deployed in the policy response to disinformation online, it is worth considering a hypothetical statutory rule of its form, namely: *firms must implement effective and proportionate measures to limit the virality of disinformation on their services*. Notably, the qualitative and behavioural nature of the principles-based rule form allows for far greater

5. The German NetzDG (2017) and the EU Terrorist Content Online regulation (COM (2018) 640 final) are two cases in point of this trend.

flexibility in the types of measures that demonstrate compliance. Relevant OSPs could, depending on their specific context, satisfy this rule by implementing: an automated mechanism to identify content that meets a standard of virality, allowing for fast-track review; a content de-ranking policy for identified disinformation; or, specific OCR system design tweaks. Another such principle-based rule to address disinformation could be: *firms must take steps to enhance the visibility of factual information in OCR systems*. As with the previous example, the desired regulatory outcome here could be achieved in a number of different ways, as evidenced by the various voluntary efforts of platforms such as Facebook and YouTube to that specific end to date.

Again, the move towards principles-based rule-structuring would likely entail new loci of regulatory intervention. Principles-based rules invite compliance measures that manifest *earlier* in the product cycle—such as in the design and operation of OCR systems—as these interventions are likely to be far more influential in responding to the rules' behavioural and purposeful requirements. For instance, the effort to limit virality of disinformation is likely to be far better satisfied by measures that correct for the very design flaws in OCR systems *that give rise to said virality*, rather than an intervention late in the product cycle that focuses on removal speed of notified content. [6]

Again, while not ostensibly concerned with the regulation of disinformation online, the DSA signals a new embrace of the principles-based approach to rule form in the online content domain. For instance, the risk mitigation measures that firms are expected to take under article 27 are to be 'reasonable, proportionate, and effective' while also being 'tailored' to specific risks. In addition, firms operating OCR systems must provide service users with 'clear, accessible, and easily comprehensible' information regarding the curative role of these systems (art. 29). The DSA also seeks to anticipate the *compliance challenges* that are likely to accompany the shift to principles-based rule forms, most notably in its commitment to 'support and promote the development and implementation of voluntary industry standards' (art. 34) and its provisions on future Codes of Conduct that will aim to elaborate on, and give specific meaning to, the generalised rules (art. 35). It is important to note that while the DSA will not set out specific principles-based rules for addressing *disinformation* (such as those used in the examples previously), it will nonetheless provide the novel regulatory architecture *within which* such rules can be developed and implemented. Indeed, the European Commission has already

6. Of course, this assumes that companies and those overseeing them are committed to implementing the principles-based approach in good faith, an assumption that is interrogated in Section 5.

suggested that the principles-based EU Code of Practice on Disinformation will, in the future, be subsumed under the legal architecture of the DSA's Code of Conduct provisions (European Commission, 2020 §4.2).

3.4 The 'responsibilisation' of internal management

The third and final feature of financial services regulation that is relevant for our purposes is the explicit reliance on regulated entities in the execution of regulatory objectives. Firms are *themselves* expected to take primary responsibility for operationalising generalised rules in their own internal compliance programme and devising the means by which the rules' objectives are best achieved. Examples of this philosophy can be found across the financial services legislative *acquis*. For instance, the Prudential Requirements directive (Directive 2013/36/EU) directs national regulators to 'ensure oversight *by the [firm's] management body*, promote a sound risk *culture at all levels* of [...] firms and [...] monitor the adequacy of *internal governance arrangements'* (Ibid, art. 54, my emphasis). In the theory, this strategy of empowering and relying on firms to achieve regulatory objectives is typically referred to as 'management-based' or 'meta-' regulation (Black, 2012; Coglianese & Lazer, 2003). The management-based theory of regulation aims to incentivise commercial prudence at the point where firms are *contemplating* business decisions and practices, rather than in outputs.

When applied to the disinformation context, rather than being commanded and controlled through specific directives, relevant OSPs would bear primary responsibility for identifying and mitigating the risks that their commercial practices pose to the achievement of regulatory objectives regarding disinformation. Practically-speaking, it would likely materialise in an increased focus on formalised impact assessments that are tailored to disinformation-rated risks; systemic documentation by firms of their internal operational processes; and, an organisational restructuring that gives more prominence to internal compliance functions (e.g., the creation of a Chief Risk Officer; embedding compliance staff in product teams; etc.). As with the risk-based and principles-based approaches, the management-based approach will likely shift the focus of compliance measures towards OCR systems, as it forces firms to scrutinise how their *business practices* are likely to contribute to the spread and impact of disinformation. Ultimately, the specific compliance strategies that firms pursue under the management-based approach could be subject to varying degrees of supervision, depending on the levels of trust between firms on the one hand, and the public and policymakers on the other. [7]

7. For instance, firms could be obliged to submit their chosen compliance strategy to regulatory au-

Again, we see some flavour of the management-based approach in the draft DSA, with *firms themselves* given the responsibility to identify, evaluate, and manage the risks that *they* face (arts. 26, 27); develop 'audit implementation report[s]' in response to recommendations of third-party auditors who monitor the law's enforcement (art. 28); and to appoint compliance officers who 'shall directly report to the highest management level of the platform' (art. 32).

At this point we have outlined how, by deploying certain foundational theories of financial services regulation, we could develop a novel policy approach to disinformation online. We have also seen how these theories are already starting to enter the EU content regulation paradigm, through the provisions of the DSA. Equipped with this context, we can now turn to the crucial question of whether a regulatory transformation on this basis could *meaningfully address* the identified shortcomings of the contemporary regulatory approach to disinformation online.

Section 4. Improvements on the status quo

4.1 The appeal of content agnosticism

This project is motivated by the belief that our contemporary approach to disinformation online systematically engenders intolerable interferences with individuals' freedom of expression, owing to its reliance on performance- and technological-based theories of regulation. 'Compliance success' under this 'content-centric' approach means *more* takedowns and *more* filtering, a state of affairs that does not apply well to a type of content the harm of which is dependent on a variety of contextual factors.

By borrowing regulatory theories from the financial services sector we could alleviate this problem to a considerable degree, simply because this approach would tend towards regulatory interventions that locate themselves '*upstream*' in the product cycle. For instance, the risk-based approach gives recognition to the fact that firms' business *practices and processes*—in particular, OCR systems—can contribute significantly to disinformation-related harms. As such, the approach aims at interventions *vis-à-vis* those very practices and processes. It is, as such, content *agnostic*—firms would not be explicitly directed to take action against specific pieces of online content and in the majority of cases there would not be an expectation that objectionable content be removed from the public domain (MacCarthy, 2020). [8] It simply requires firms adopt a more prudential approach in their business prac-

thorities for prior review; be obliged to keep formal compliance strategies 'on file' for ex post review; or simply be subject to ad hoc 'spot checks' by regulatory authorities.

tices, and particularly with respect to how they *commercially engage* with the content that users upload to their services (e.g., how they amplify, target, and present it to new audiences). Indeed, 'success' under my envisaged approach could occur even if a firm filters or renders inaccessible *no disinformation at all*, but simply designs its OCR system such that content that is likely to be disinformation is not purposefully *amplified* and *micro-targeted* to those users for whom it may cause untenable harm. Ultimately, the risk-based approach aims to regulate *firms,* not the *users* of those firms' products.

4.2 Reorienting regulation around the harmful practices

Section 2 also identified how the regulatory theories underpinning the contemporary approach eschew interventions that target the business practices that inform the harm in disinformation, namely the content's promotion, targeting, and presentation through OCR systems.

Fortunately we have now seen how the theories underpinning financial services regulation unlock the means by which we can effectively intervene *vis-à-vis* said commercial practices. For instance, the risk-based approach renders OSPs that operate OCR systems accountable for ensuring the systems *themselves* are designed and operated with diligence, rather than simply obliging the firm to 'clean up' the bad outcomes that those systems give rise to. By reorienting interventions towards *commercial practices* and away from the substantive *content*, the risk-based approach acts on our recognition that the policy problem of disinformation is dependent on the various contextual factors that inform its consumption; many of which are controlled by firms. Put another way, the risk-based approach allows us to address one of the key causal factors of disinformation as a *policy problem*, rather than merely its *symptoms*. [9]

Furthermore, the principle-based approach to rule-structuring—compared to the traditional 'bright-line' or 'prescriptive' forms—is likely to ensure that compliance interventions remain meaningful and focused on OCR systems through time, even

8. Admittedly, there may be some discrete contexts where it may still be wise to define risk mitigation efforts in terms of content filtering and removal, particularly where specific disinformation is likely to pose serious and imminent harm to a sufficiently large number of people (e.g. disinformation encouraging the drinking of bleach as a miracle cure for COVID-19 just after a speech by an influential public figure that endorses the claim). In any case, I suspect these cases to be the exception rather than the norm.

9. Again, this is not to suggest that the problem of disinformation can be reduced to and solved by addressing interventions towards OCR systems alone. Rather, it seeks to give due regard to, and provide an effective response to, the considerable role played by such systems in compounding the broader problem.

in the face of rapid technological and operational change. Principle-based rules set out—in *qualitative* and *behavioural* terms—how firms are expected to act and what outcome they are expected to achieve. Intuitively, legal mandates of this form are more difficult to circumvent, as their qualitative and object-orientated nature means it is easier to assess whether a given action on the part of firms amounts to a *genuine* effort to achieve the regulatory goal. This approach can thus help us avoid the regulatory phenomenon of 'hitting the target but missing the point' (Black, 2010, p. 7). As an example, in section 3.3 I suggested that a potential principle-based rule could take the form that *firms must implement effective and proportionate measures to limit the virality of disinformation on their services*. While the regulatory objective of this rule could be satisfied by several unique approaches, its qualitative and object-orientated nature allows an observer to make certain baseline assessments of what *meaningful* compliance looks like. In this case, the pathology of online content virality and the fact that it is largely a function of the design and operation of OCR systems means that a compliance programme that ignores OCR systems is unlikely to be a good-faith effort. Moreover, those same rule-form characteristics mean OCR systems can remain the focal point of intervention through time, even if the technological and operational features of those systems evolve.

Finally, I discussed how the commercial incentives that underpin the design and operation of OCR systems are a key reason why those systems can unintentionally contribute to the spread and impact of disinformation online. While it is difficult to accurately *measure* this divergence between public and private incentives in the OCR sector, the anecdotal evidence that we referenced in Section 2 suggests that it is pronounced. Fortunately, the management-based approach can help correct for this tendency, by ensuring *firms* take primary responsibility for their social outputs and by facilitating the 'internalisation' of commercial prudence in the long-run. Experience in the financial services sector and beyond suggests that private stakeholders are more likely to view compliance measures and processes as worthy of adherence if they are the ones devising those specific measures (Black, 2012; Coglianese & Nash, 2006; Coglianese & Lazer, 2003). The responsibilisation of internal management under this regulatory theory thereby engenders a sense of 'ownership' in the pursuit of regulatory objectives (Gunningham & Sinclair, 2017). Given the breath of sectors where this phenomenon has been witnessed, it is reasonable to believe that a similar organisational psychology is likely to hold in the OCR sector, and thus the management-based approach can contribute to a narrowing of the sectoral incentives gap (Coglianese & Lazar, 2003).

Section 5. Appreciating the challenges

5.1 Lingering rights concerns

Admittedly, concerns regarding unjustified interferences with fundamental rights remain pertinent. While this approach aims to address the most *acute* interference that is foreseeable in the content regulation domain—namely the excessive blocking and removal of legitimate expression—novel concerns will come to the fore under this envisaged model. For instance, there are important questions as to whether down-ranking or de-ranking of certain content amounts to an excessive interference with an individual's freedom of expression (e.g. 'shadow-banning'), even if that content remains somewhat discoverable on the platform. Similarly, how are we to react when upstream compliance measures that aim at improving OCR systems are discovered to result in disparate impact on important public interest expression and group perspectives (e.g. an OCR system stops or significantly reduces its promotion of 'Black Lives Matter' activist content)?

In addition, one cannot ignore the critique that to deploy financial services regulatory theory to our domain is to do no more than legitimise the problem of 'privatised enforcement', the phenomenon whereby OSPs take the place of legislative and judicial authorities as the primary (and often sole) arbitrators of online speech, without the traditional legal safeguards that regulate state-driven interferences with individuals' fundamental rights. The risks to fundamental rights—notably the right to receive and impart information; the right to privacy and data protection; and various due process rights—that can arise from the phenomenon of privatised enforcement are well established in the literature (Kuczerawy, 2017; Belli & Venturini, 2016; Angelopoulos et al., 2016).

Unfortunately, the problems associated with privatised enforcement *could* manifest in our suggested approach, particularly if it is ultimately unable to prevent firms from expressing their compliance in the 'content-centric' manner we see today. Indeed, it is reasonable to assume that some firms in the OCR sector will simply aim at 'achieving compliance' at the least cost to the business—that is, by building-on existing trust & safety programmes while ring-fencing their business practices and processes from any meaningful alteration. In addition, even where we assume *good intent* on the part of firms, it is possible that in practice some firms' compliance functions will be unable to meet the rigorous expectations of this approach and may thus adopt excessively conservative or 'industry-standard' compliance strategies. The suggested approach would place primary responsibility on firms to identify and manage the disinformation-related risks that they face, while at every

stage ensuring that they are acting towards the regulatory objectives as established in the generalised rulebook. That is no easy task, and as a result some well-intentioned firms may retreat into 'content-centric' approaches to minimise compliance risk. Indeed, rather than ushering in a modern era of rights-protective responsive regulation, this approach could quickly backslide into one that incentivises *more* blocking and removal, with *less* public oversight.

These risks cannot be underappreciated, and should be at the fore of policymakers' considerations when contemplating regulatory interventions on the basis of the proposed regulatory approach. That said, given that many of the aforementioned risks are likely to manifest in the *implementation* of the regulatory approach, it may be possible to anticipate and manage them through thoughtful legislative crafting, rigorous oversight and monitoring, and a focus on safeguards-by-design. For instance, an obvious means to protect against the novel risks to freedom of expression that may arise through 'upstream' interventions aimed at OCR systems would be through obligations on regulated firms to develop public-facing policies on down- and de-ranking content; mandating third party algorithmic audits and impact assessments that monitor for discriminatory outcomes or disparate impact on certain groups' expression; and mandating disclosure to public authorities and researchers of the specific content that has been impacted by content moderation practices. Furthermore, to address the risk of backsliding into rights-interfering content-centric approaches and opening the door to the worst excesses of 'privatised enforcement', policymakers could consider mechanisms for ensuring rolling behavioural oversight of regulated entities, and supports for firms to translate general statutory obligations into tailored compliance measures (e.g., through secondary legislation; co-regulatory codes of conduct; etc).

5.2 The 'riskification' of policy

While the endeavour to place risk at the centre of the content regulation paradigm would be in keeping with a broader trend within digital policy, and indeed public policy more generally, the 'riskification' of regulation is not without its problems (Beck, 1992).

First, as Böröcz (2016) notes in the context of the General Data Protection Regulation (GDPR) risk-based approach, fundamental rights are principally products of the legal domain, and enjoy their own distinct meaning and epistemic coherence. Risk management as a discipline is something altogether different, and bases itself predominantly within a techno-scientific epistemology. As such, the two domains do not always mesh together in a coherent manner, and it can be challeng-

ing, if not impossible, to understand fundamental rights in terms of risk and understand how different risk factors can promote or limit the enjoyment of certain fundamental rights. Indeed, as van Dijk et al. (2016, p. 289) observe, when rights and risks are conflated 'the meaning of both are changed into something that could hardly be predicted in advance'. This difficulty is even more pronounced when our public policy interest is to manage risks that may only manifest on a *societal* level (e.g. the risk to electoral integrity and democratic discourse). In these contexts, the link between a certain business practice or product and risks to fundamental rights may not be obviously expressible in quantifiable terms, and fundamental rights may not even be the most appropriate *yardstick* against which to measure the harm in question (McKay & Tenove, 2020).

Second, risk management, like all techno-scientific disciplines, is value-laden. Value judgements are made when one chooses *what* risks to assess, *how* to assess them, and ultimately, how to *manage* then. As acclaimed risk theorist Paul Slovic has noted, 'defining risk is an exercise of power' (Buni & Chemaly, 2020, §2). Slovic's observation is particularly concerning in our context, given that the centres of power in the technology sector are overwhelmingly male, white, and located in the global North (Harrison, 2019). As such, it is likely that—independent oversight notwithstanding—in the endeavour to assess, evaluate, and manage risks related to their service, firms will overlook serious risks that pertain to certain groups or adopt a risk-appetite posture that runs counter to reasonable expectations of the public and policymakers. Buni & Chemaly (2020) document numerous case studies that attest to how bias and a systemic diversity problem has shaped a sub-optimum risk management culture in the tech sector, encompassing everything from market risk ('what risks do the political and social conditions in the region we're deploying give rise to?'); to use-case risk ('what bad outcomes to certain users and groups are likely to arise from use of our product?'); to trust and safety risk ('in managing a known risk, what new risks are we likely to stimulate?').

EU lawmakers would do well to heed these learnings from other sectors when considering the risk-related provisions of the DSA.

5.3 The thorny question of oversight

It has already been noted that the shift towards a regulatory model for disinformation online characterised by risk-based, principles-based, and management-based approaches is likely to place considerable compliance challenges on firms, and if poorly implemented, could pave the way for the interferences with rights (e.g. freedom of expression) that the proposal aims to mitigate. It is in recognition of

similar inherent 'implementation risks' that financial services policymakers have opted to include independent regulatory agencies in their regulatory architecture. Bodies like the European Banking Authority and the UK Prudential Regulation Authority act as regulatory fulcrums, giving meaning and reality to policymakers' political aspirations. It is thus unsurprising that, given its tentative endorsement of the types of regulatory theories that warrant agency-led oversight, the DSA places a similar oversight model at the centre of its proposed regulatory architecture (art. 38; art. 51).

Yet things are rarely simple with respect to agency-led oversight, especially in the context of online content regulation. For a start, instituting new regulatory authorities brings its own risks. First, there is the obvious risk that such bodies can be 'captured' by firms in the market. Viewed cynically, a firm needs not actually *meet* regulatory objectives, it must simply *convince the regulator* that it is doing so. What regulators see is not necessarily a compliance programme that a firm has implemented to achieve the regulatory objectives, but rather, a *representation* or *idealisation* of such a programme. Indeed, a critique of regulatory oversight in the financial services sector is that firms focus less on managing their own compliance processes, and more on *managing the regulator* (Black, 2012, p. 1047; Anderson 1982). Second, agency-led oversight of the (social) media sector has traditionally been viewed with skepticism in many jurisdictions, and my suggested approach could evoke images of a 'ministry of truth' and state censorship. That critique is not without merit, and there is a real risk that regulatory bodies themselves may 'backslide' into an approach that views compliance success under this paradigm in terms of ever more content removal in ever shorter periods of time. Indeed, bestowing new powers on regulatory authorities can be a dangerous exercise when those regulators' mission and intent is subject to doubt (Article 19, 2021).

While these concerns are very real, they may not be *immutable*. Regulatory authorities can be resilient against corporate and cognitive capture if the institutional set-up and broader political context boasts three features: first, a high-degree of transparency in public affairs, so corporate influence can be kept in check; second, technical expertise and investigative resources within oversight authorities, so regulators can make diligent and effective assessments as to firms' compliance efforts; and ultimately, strong government support for their mission, so said regulators can execute their mandate in the face of market and media pressure. [10] Of course, said

10. As a case in point, not all financial services regulatory authorities failed in executing their oversight functions in the lead-up to the 2008 global crisis, and those that fared relatively well appear to have shared the aforementioned qualities (Black, 2014).

government support will do more harm than good if the government in question seeks to co-opt agency-led regulation to suppress fundamental rights. In these situations, and where the jurisdiction in question is the site of systematic interference with the rule of law from state institutions, agency-led oversight—and perhaps the novel regulatory approach *tout court*—should not be considered in the first instance.

6. Conclusion – from theory to practice

At this point I have identified key regulatory theories underpinning financial services regulation in Europe, and articulated how they could be transposed to address disinformation online. I have illustrated how transposition would serve as an improvement on the regulatory status quo, in that it would allow us to locate regulatory interventions at OCR systems while mitigating impacts on freedom of expression. I have also articulated the potential implementation risks associated with this approach, and proffered tentative solutions for how they might be assuaged. What then, is left for us to do? Arguably, this project has achieved its main objective—that is, to verify that the envisaged regulatory transposition holds promise and should be considered and explored by European policymakers and the policy community. In that context, I will conclude by marking out a path to guide the next stages of this endeavour—what needs to happen to allow us to realise this promising alternative approach.

First, we must define the *precise* principle-based rules that we believe can best address disinformation in OCR systems. In section 3 I deployed examples that relate to virality and authoritative sources. Of course, these are not the only possible principles-based rules, and there are a whole host of others that could complement or replace them. Second, my approach places risk at its core—as both the grounding for regulatory legitimacy and the metric by which firms' obligations ought to be defined. I have already noted that for a substantive framework to take shape, we must first develop a deeper understanding of risk in this context. The next stage must develop that conception of risk, as it is a crucial foundational pillar upon which we can consider what types of efforts the OSP sector should be undertaking to address disinformation in its various forms. Third, good policy depends on rich data, and to develop the best principle-based rules and an appropriate conception of risk we need better insight into how disinformation manifests online. Indeed, we still have little *systematic* knowledge of how precisely disinformation spreads through OCR systems and the true contribution of firms' commercial practices to it (Leersson, 2000; Haim & Nienierza, 2019; Ingram, 2019). That

needs to be addressed, and urgently.

Fortunately, the draft DSA offers a crucial opportunity to advance in each of these endeavours. Most notably, the draft law's transparency requirements—a blend of third-party auditing requirements; investigatory powers for regulators; and the various obligations to provide transparency to the research community and to the public—can address the asymmetry of information that stifles policy work in this domain. Moreover, as noted in section 3.2, the draft DSA provides a rudimentary conception of risk in the online content domain. Scrutinising, problematising, and ultimately improving this conception will be a necessary challenge and one which can inform future policy deliberations around disinformation online. Finally, while the DSA is ultimately unlikely to set out principles-based rules for addressing disinformation, it will provide the regulatory architecture *within which* such rules can be developed and implemented in the future.

In closing, I am under no illusions about the challenges inherent in pursuing this project through its next stages. I have taken merely the first step with this article. Yet I draw comfort from the fact that we have completed perhaps the most important step, as all practical regulation is underpinned by an initial theoretical hypothesis.

References

Anderson, J. (1982). The public utility commission of texas: A case of capture or rapture? *Review of Policy Research*, *1*(3). https://doi.org/10.1111/j.1541-1338.1982.tb00453.x

Angelopoulos, C., Brody, A., Hins, W., Hugenholtz, B., Leerssen, P., Margoni, T., McGonagle, T., van Daalen, O., & van Hoboken, J. (2016). *Study of fundamental rights limitations for online enforcement through self-regulation* [Report]. Institute for Information Law (IViR), University of Amsterdam. https://hdl.handle.net/1887/45869

Article 19. (2021). *At a glance: Does the EU digital services act protect freedom of expression?* Article. https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/

Ayres, I., & Braithwaite, J. (1992). *Responsive regulation: Transcending the deregulation debate*. Oxford University Press.

Balkin, J. (2018). *Fixing social media's grand bargain* (Paper No. 1814; Aegis Series). Hoover Institution. https://www.hoover.org/sites/default/files/research/docs/balkin_webreadypdf.pdf

Beck, U. (1992). *Risk society: Towards a new modernity*. SAGE Publications.

Belli, L., & Venturini, J. (2016). Private ordering and the rise of terms of service as cyber-regulation. *Internet Policy Review*, *5*(4). https://doi.org/10.14763/2016.4.441

Bergen, M. (2019). YouTube executives ignored warnings, letting toxic videos run rampant. *Bloomberg*. https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Black, J. (2008). *Forms and paradoxes of principles-based regulation* (Working Paper No. 13/2008; LSE Law, Society and Economy). London School of Economics and Political Science. https://www.lse.ac.uk/law/working-paper-series/2007-08/WPS2008-13-Black.pdf

Black, J. (2010). *The rise, fall, and fate of principles-based regulation* (Working Paper No. 17/2010; LSE Law, Society and Economy). London School of Economics and Political Science. https://core.ac.uk/download/pdf/17332.pdf

Black, J. (2012). Paradoxes and failures: "New Governance" techniques and the financial crisis. *The Modern Law Review*, *75*(6). https://doi.org/10.1111/j.1468-2230.2012.00936.x

Böröcz, I. (2016). Risk to the right to the protection of personal data: An analysis through the lenses of Hermagoras. *European Data Protection Law Review*, *2*(4). https://doi.org/10.21552/EDPL/2016/4/6

Buni, C., & Chemaly, S. (2020). The risk makers [Medium Post]. *OneZero*. https://onezero.medium.com/the-risk-makers-720093d41f01

Cobbe, J., & Singh, J. (2019). Regulating recommending: Motivations, considerations, and principles. *European Journal of Law and Technology*, *10*(3). https://ejlt.org/index.php/ejlt/article/view/686

Coglianese, C., & Lazer, D. (2003). Management-based regulation: Prescribing private management to achieve public goals. *Law and Society Review*, *37*(4). https://doi.org/10.1046/j.0023-9216.2003.03703001.x

Coglianese, C., & Nash, J. (2006). *Leveraging the private sector: Management-based strategies for improving environmental performance*. Routledge. https://doi.org/10.4324/9781936331444

Dijk, N. (2016). A risk to a right? Beyond data protection risk assessments. *Computer Law & Security Review*, *32*(2).

Engstrom, E., & Feamster, M. (2017). *The limits of filtering: A look at the functionality & shortcomings of content detection tools*. Engine. https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf

Francois, C. (2019). *Actors, behaviours, content: A disinformation ABC* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression Series). Annenberg Public Policy Center, University of Pennsylvania; Annenberg Foundation Trust, Sunnylands; Institute for Information Law, University of Amsterdam. https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf

Gary, J., & Soltani, A. (2019). *First things first: Online Advertising practices and their effects on platform speech* (Free Speech Futures) [Essay]. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech

Gillespie, T. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1512

Graef, I., & Van Berlo, S. (2020). Towards smarter regulation in the areas of competition, data protection, and consumer law: Why greater power should come with greater responsibility. *European Journal of Risk Regulation*, *25*(1). https://doi.org/10.1017/err.2020.92

Gunningham, N., & Sinclair, D. (2017). Trust, culture and the limits of management-based regulation: Lessons from the mining industry. In *Regulatory theory*. ANU Press. https://doi.org/10.22459/RT.02.2017.40

Haim, M., & Nienierza, A. (2019). Computational observation: Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. *Computational Communication Research*, *1*(1). https://doi.org/10.5117/CCR2019.1.004.HAIM

Haines, E. (2019). Manipulation machines: How disinformation campaigns suppress the Black vote. *Columbia Journalism Review*. https://www.cjr.org/special_report/black-misinformation-russia.php

Hao, K. (2021). How Facebook got addicted to spreading misinformation. *MIT Technology Review*. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Harrison, S. (2019). Five years of tech diversity reports – and little progress. *WIRED*. https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/

Helberger, N. (2020). The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. *Digital Journalism*, *8*(3). https://doi.org/10.1080/21670811.2020.1773888

Horwitz, J., & Seetharaman, D. (2020). Facebook executives shut down efforts to make the site less divisive. *Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Ingram, M. (2019). Silicon valley's stonewalling. *Columbia Journalism Review*. https://www.cjr.org/special_report/silicon-valley-cambridge-analytica.php

International Organisation for Standardisation. (2018). *ISO 31000 risk management – guidelines*. https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en

Keller, D. (2019). *Dolphins in the net: Internet content filters and the advocate general's Glawischnig-Piesczek v. Facebook Ireland opinion* [White Paper]. Stanford Center for Internet and Society, Stanford Law School. https://stanford.io/3kOfrKv

Kuczerawy, A. (2017). The power of positive thinking: Intermediary liability and the effective enjoyment of the right to freedom of expression. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law*, *8*(3). https://nbn-resolving.org/urn:nbn:de:0009-29-46232

Leerssen, P. (2020). The soapbox as a blackbox: Regulating transparency in social media recommender systems. *European Journal of Law and Technology*, *11*(2). http://www.ejlt.org/index.php/ejlt/article/view/786

MacCarthy, M. (2020). *Transparency requirements for digital society media platforms: Recommendations for policymakers and industry* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression Series) [Working Paper]. Annenberg Public Policy Center, University of Pennsylvania; Annenberg Foundation Trust, Sunnylands; Institute for Information Law, University of Amsterdam. https://www.ivir.nl/publicaties/download/Transparency_MacCarthy_Feb_2020.pdf

Macrotrends. (2021). *Revenue comparison through time – Facebook, Twitter*. Macrotrends. https://www.macrotrends.net/stocks/stock-comparison?s=revenue&axis=single&comp=FB:TWTR

Maréchal, N., & Roberts Biddle, E. (2020). *It's not just the content, it's the business model: Democracy's online speech challenge* [Report]. Open Technology Institute, New America. https://www.newameric

a.org/oti/reports/its-not-just-content-its-business-model/

McKay, S., & Tenove, C. (2020). Disinformation as a threat to deliberative democracy. *Political Research Quarterly*. https://doi.org/10.1177/1065912920938143

Shmargad, Y., & Klar, S. (2020). Sorting the news: How ranking by popularity polarizes our politics. *Political Communication*, *37*(3). https://doi.org/10.1080/10584609.2020.1713267

Singh, S. (2019). *'Rising through the ranks: How algorithms rank and curate content in search results and on news feeds* [Report]. Open Technology Institute, New America. https://d1y8sb8igg2f8e.cloudfront.net/documents/Rising_Through_the_Ranks_2019-10-21_134810.pdf

Solsman, J. (2018, January). YouTube's AI is the puppet master over most of what you watch. *CNET*. https://www.cnet.com/news/youtube-ces-2018-neal-mohan/

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408

van Dijk, N., Gellert, R., & Rommetveit, K. (2016). A Risk to a Right? Beyond Data Protection Risk Assessments. *Computer Law & Security Review*, *32*(2), 286–306. https://doi.org/10.1016/j.clsr.2015.12.017

Wagner, B., Kübler, J., Kuklis, L., & Ferro, C. (2021). *Auditing big tech: Combating disinformation with reliable transparency* [Report]. Enabling Digital Rights and Governance. https://enabling-digital.eu/wp-content/uploads/2021/02/Auditing_big_tech_Final.pdf

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Towards an interdisciplinary framework for research and policymaking* (Report DGI(2017)09). Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Woods, L., & Perrin, W. (2019). *Online harm reduction – a statutory duty of care and regulator* [Report]. Carnegie UK Trust. http://repository.essex.ac.uk/25261/1/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

Legislation and governmental documents

Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC Text with EEA relevance, 32013L0036, EP, CONSIL, OJ L 176 (2013). http://data.europa.eu/eli/dir/2013/36/oj/eng

European Commission. (2018). *Final results of the Eurobarometer on fake news and online disinformation* [Report]. European Commission. https://ec.europa.eu/digital-single-market/en/news/final-results-eurobarometer-fake-news-and-online-disinformation

European Commission. (2020). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions—On the European democracy action plan. COM(2020) 790 final*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz—NetzDG), [('NetzDG, 2017)'] BGBI. I S. 3352, (2017). https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html

House of Commons, Treasury Committee. (2009). *Banking Crisis: Regulation and supervision* (Report No. 14; Session 2008–09). House of Commons. https://publications.parliament.uk/pa/cm200809/cmselect/cmtreasy/767/767.pdf

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. (2020). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0825&from=en

Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online, COM(2018) 640 final. (2018). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52018PC0640

Prudential Regulation Authority. (2014). *PRA rulebook: Fundamental rules instrument 2014*. http://www.prarulebook.co.uk/rulebook/Media/Get/308c054e-fae4-4e41-90dd-4826c139e2ae/PRA_2014_17/pdf

Published by

Alexander von Humboldt
INSTITUTE FOR INTERNET
AND SOCIETY

in cooperation with

centre
– internet
et société