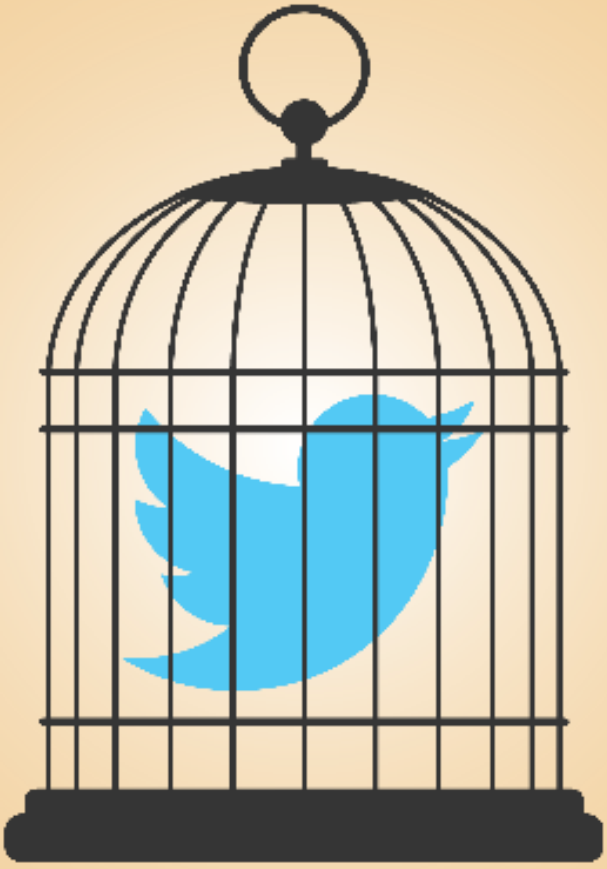
Free Speech Union briefing

Who Watches the Watchmen?

Ofcom, free speech and the Online Safety Bill

Frederick Attenborough

April 2022



FSU
Free Speech Union

About the author

Dr. Frederick Attenborough was previously Lecturer in Communication and Media Studies at Loughborough University (2009-15) and Programme Leader and Senior Lecturer in Sociology at Bishop Grosseteste University, Lincoln (2015-21). Currently based at the University of Lincoln, he is also the Communications Officer for the Free Speech Union. Since 2020 he has been a regular contributor to Toby Young's *Daily Sceptic*, and many of those pieces are now collected together on his Substack account, MyBodyThisPaperThisFire.

Contents

Introduction

The latest version of the Online Safety Bill was introduced in Parliament on 16 March 2022[1]. Subject to scrutiny, debate and some no doubt minor amendments it is likely that this iteration of the proposed legislation will be passed and made into law. According to the inimitably boosterish rhetoric so beloved of Boris Johnson's administration, the UK isn't about to be fobbed off with any old journeyman, making-up-the-numbers online safety laws, either. No, from now on we're all going to be lucky enough to have our online activities policed, chivvied and curtailed by what are in fact "*world-leading* online safety laws".[2]

World-leading on whose criteria, one wonders?

Certainly it's true that the primary aim of the Bill's (no doubt suitably *world-leading*) regulatory framework will be to remove "illegal online content" and "priority content that is harmful to adults", as well as content that is harmful to children.[3] What we shouldn't lose sight of in the Government's giddy rush to celebrate the creation of their "safer online environment", however, is the fact that the Bill professes to having a secondary aim; an aim that is, in some ways, precipitated by the first. Early on in the law-making process the Government seem to have twigged that you can't go about arbitrarily decreeing what can or can't be uttered, filmed, drawn, written, seen, debated, analysed, critiqued, posted, uploaded or Tweeted online, and then expect to be able to strut about on the global stage like some sort of Lockean paragon of liberal democratic virtue without one or two of the bawdier citizenry down in the cheap seats at the back of the stalls starting up with a spot of the old slow handclap treatment. Perhaps that's why the latest iteration of the Bill pays some grudging, and frankly rather desultory, attention to its secondary aim, namely, ensuring that the proposed regulatory system will "strengthen people's rights to express themselves freely online and ensure social media companies are not removing legal free speech".[4]

Not that it's particularly easy to locate that attention. Approaching the vast, grey-stoned and crenellated sham of this Bill's legal architecture through the overly ornate wrought-iron gates out front, we find ourselves caught

in the glare of its primary aim: *providing individuals with a secure, stable and essentially sanitised online environment*. It's only when we creep around the back that we're able to discern the outline of the Bill's secondary aim, lurking shamefacedly in the shadows behind the bin-store just opposite the servant's entrance: *allowing for the risk and the uncertainty of a public sphere in which freedom of speech is not just tolerated but actively cultivated.* Ostensibly, these aims are incompatible… and yet the Bill claims to be forcing them into a position of compatibility. It's true, of course, that much of what passes for 'politics' in the post-Enlightenment Anglosphere is little more than a collective, emotionally fraught attempt to do just that, with a variegated assortment of charismatic gargoyles all dashing about trying to cadge a soapbox from which to announce that they, and they alone, have found the best way, the *only* way, the *Net Zero* way, the *Third Way*, to force security into compatibility with freedom, becoming into compatibility with being, the state with individuals, certainty with risks, social order with innovations, collective responsibilities with rights. And yet it is, nonetheless, a *little* unusual for an online regulatory system to find itself tasked with resolving the sorts of epistemological dichotomies that the Montesquieus, de Tocquevilles, Hegels, Hayeks, Schumpeters, Nietzsches, Polanyis, Heideggers, Arendts, Sartres and Foucaults of this world have all hitherto found so intractable.

So will the regulatory system proposed by the Online Safety Bill prove at all capable of fulfilling what is, essentially, an ethico-philosophical task? Will it possess the necessary intellectual dexterity, the necessary political commitment, not just to hold these apparently incompatible principles in tension, but to champion them both and to do so at the same time? Will it come into being armed with regulatory tools and mechanisms potent enough to hold multi-billion-dollar online providers to account for the adequacy (or otherwise) of the balance their online content moderation systems establish between collective security *and* individual freedom? Or might the Bill's references to "strengthen[ing] people's rights to express themselves freely online" prove to be little more than well-intentioned makeweights; a form of linguistic ballast that rolls rather nicely off one's tongue in the rarefied atmosphere of the Houses of Parliament, say, or perhaps while sitting freshly varnished on a breakfast TV sofa and incanting one's vacuous, learnt-by-rote soundbites; linguistic ballast that, in practice, out there in Silicon Valley's digital swamp, might all too quickly come to be discharged, cut adrift, forgotten about, once the proposed regulator's staff realise they are laughably out of their depth, drowning in algorithms they cannot hope to understand, and surrounded by sharks that will

continue to tear the First Amendment to pieces, whether these ‘world-leaders’ want them to or not?

It is to these and other, similar questions that this paper addresses itself.

1. Ofcom – regulator in legislation, regulator in practice?

The Government originally set out its intention to appoint Ofcom as the regulator in legislation via the Online Harms White Paper (April 2019),[5] before the full response to the White Paper consultation (December 2020) confirmed that this would indeed be the case.[6] It's therefore critical for anyone with an interest in defending free speech to understand the scale and scope of Ofcom's new role. In the legal ecosystem proposed by the Online Safety Bill,[7] Ofcom will be attempting to do for online platforms what it has been doing for some time now in the case of broadcasters, namely, "tackling harmful content and [*rather more importantly for us*] protecting freedom of speech".[8] How will it attempt that protection? Via the issuance of statutory codes of practice requiring online service providers to establish free speech 'safeguards'.

At first glance that might seem rather reassuring – so much so, in fact, there's a danger of being lulled into a false sense of security; of coming away from a brief perusal of the draft Bill with the feeling that these 'codes' will allow Ofcom to establish the right balance between society's right to uphold certain normative standards and each individual member of that society's right to freedom of speech. A useful corrective is to move a little further downstream of the legislative process and consider what's likely to happen once the Bill passes into law, the variegated assortment of Parliamentarians, academics, campaign groups and celebrities currently clustering around it have all buzzed off elsewhere and Ofcom is left alone to grapple with the messy, day-to-day practicalities of enforcing its safeguards.[9]

Oddly enough, that's exactly the proleptic journey Ofcom's Chief Executive, Dame Melanie Dawes, was asked to make during her recent appearance before the Draft Online Safety Bill Joint Committee. Did she feel Ofcom would have what it needed to act and act quickly if ever it were required to hold the big tech giants to account? "Broadly, yes," she said.[10] It's hardly surprising her answer feels weighted down

with all manner of unspoken caveats and qualifiers: although it is just about arguable that the regulatory mechanisms included within the draft Bill are robust enough *in principle*, there are urgent question marks regarding the level of meaningful control Ofcom will have over them *in practice*.

In part, this is an issue of capability. Even if one *were* minded to accept the Government's claim that Ofcom is an actor blessed with "organisational experience, robustness and experience of delivering challenging, high-profile remits", its ability to perform this new, action-packed role of regulator-cum-policeman without blowing up in its lines would seem far from assured.[11] It's what you might most politely describe as 'a step up' from Ofcom's current repertory theatre work, in which performances of 'regulating the country's telly and radio' and 'supervising the Royal Mail' rarely rise above the level of local village hall am-dram.[12] And for those of us who've sat through so many of their shows in the past, it's frankly rather difficult to believe that the so-called 'masters of the universe' – those global, high-tech companies possessed of multibillion-dollar research budgets, the best data scientists money can buy and proprietary, cutting-edge AI-based moderation systems that few external researchers could ever hope to access, let alone understand – will prove as susceptible to the charms of Ofcom's "organisational experience" and "robustness" as ITV, Classic FM or the bloke who pops the letters through your door most weekday mornings.[13]

Quite apart from Ofcom's capabilities, what are we to make of the company it keeps? When it comes to the tricky – indeed, some might say *impossible* – task of deciding what does or does not constitute online 'misinformation', Ofcom's door always seems to be open to Full Fact, an organisation funded by digital media companies and online service providers.[14] It's difficult to know what's more worrying about that: the fact that Ofcom regularly receives information on an important regulatory issue from an organisation funded by companies which it's going to be regulating, or the fact that Ofcom wants advice on a concept that's had such a deleterious effect on free speech.

Full Fact's funders – the Googles and the Facebooks of this world – are now openly engaged in attempts to *curate* rather than simply *facilitate* online social interaction.[15] That's why, across their various platforms, free speech finds itself locked in a dysfunctional, inverse proportional

relationship with the concept of ‘misinformation’. Whenever service users question, claim or challenge things that those companies feel shouldn’t be questioned, claimed, or challenged, their speech-acts are quickly labelled as ‘misinformation’ before being subjected to all sorts of censorship.[16] It’s in this way, slowly, over time, that accusations of ‘misinformation’ enable online service providers to seize, enclose, partition off and eventually repurpose the many wilds and common lands from which free speech would once have drawn its succour. So when an organisation like Full Fact advises Ofcom as to what constitutes ‘misinformation’ we shouldn’t forget that it is also, and at the same time, advising them on the areas of social, cultural and political life from which free speech should appropriately be withdrawn.

And what are we to make of Ofcom’s *own* attitude to free speech? Recently, for instance, Ofcom have enlarged the number of ‘protected characteristics’ in its broadcasting code from four to 18. From that point forth, broadcasters were told, ‘hate speech’ would mean: “all forms of expression which spread, incite, promote or justify hatred based on intolerance on the grounds of disability, ethnicity, social origin, sex, gender, gender reassignment, nationality, race, religion or belief, sexual orientation, colour, genetic features, language, political or any other opinion, membership of a national minority, property, birth or age”.[17] More recently, Dame Melanie Dawes weighed in on the subject of representing debates around gender and transgender, claiming that broadcasters should “steer their way through these debates without causing offence and without bringing inappropriate voices to the table”.[18] Inappropriate voices. How quaint. One imagines the Lord Chamberlain’s Office of the early-1960s issuing similar recommendations to any broadcasters toying with the idea of interviewing Joe Orton as part of a prime-time ‘modern playwrights’ series.

Perhaps there’s something else here too, though; something more contemporary, gently whispering to us in Ofcom’s actions. Let us listen more carefully. Certain categories and characteristics placed off limits during debates, deliberation and argumentation… an exhortation to broadcasters not to cause anyone any offence… a reminder that inappropriate voices should be silenced. Hmm. Is it not the case that these actions exhibit the same, indifferent, shruggy-shouldered attitude towards free speech as the online service providers? Do we not hear in these statements, these actions, the tone of a technocratic organisation that feels nothing but suspicion towards rumbustious, rollicking,

untrammelled social interaction; that looks askance at 'that sort of thing'; that is convinced that what it sees is in need of paternalist curation rather than anything quite so risky as hands-off facilitation; a regulator for which free speech, if it is anything at all, is an unnecessary extravagance, frivolous almost; a luxury to be doled out only after the people who know best have had a chance to sanitise the wilds and the common lands of public discourse; after the people who know best have made sure the little people will be able to wander freely without tripping and hurting their vulnerable little minds on any difficult ideas or arguments?

It seems a little fanciful, does it not, to imagine a regulator that isn't particularly impressed by free speech itself, ever rapping with any great ferocity the knuckles of a regulated entity that's been found to have failed in its statutory duty to safeguard it. Indeed, given that online providers will risk fines and other sanctions from Ofcom if they *don't* remove material, but will easily be able to avoid punishment for acting precipitously by demonstrating compliance with an extremely weak duty to "have regard" for free speech, there's a strong bias towards the removal of questionable-yet-perfectly-permissible-material mortared into the very architecture of this regulatory system. The Free Speech Union has argued that the Draft Online Safety Bill requires amendment to provide a positive obligation for in-scope companies to address this threat of overzealous censorship. In particular, it has argued for the imposition of two additional duties on providers, specifically:

- To take all reasonable steps to ensure that the right to freedom of expression is not unduly infringed by excessive measures taken in pursuit of the duties of care under the Bill; and
- To prepare and publish a policy setting out how this duty will be complied with.

These focused and limited measures would mean providers had to consider how the application of their other duties under the Bill would not be undertaken so excessively as to unduly infringe freedom of expression (i.e., requiring them to avoid an 'if in doubt, remove' approach, which would be highly inimical to freedom of expression). Both provisions would sit naturally in the structure of the Bill, adding to the freedom of expression duty in clause 19. As with the other duties under the Bill, they would be overseen and enforced by Ofcom and, ultimately, the

courts. In addition, a further obligation would be imposed on Ofcom to include in the risk assessment it must undertake under clause 83, an assessment of the risk of undue infringement of freedom of expression by excessive measures taken in pursuit of the duties of care. Again, this would sit neatly within the existing structure of the Bill.

A little while earlier (section 1) I likened the Online Safety Bill to a vast, grey-stoned sham of a building, and it's useful to pursue that metaphor again here now. In the absence of the Free Speech Union's amendments there's a danger that the Bill's complex legal architecture will turn out to be little more than a folly: an edifice possessed of a fine-looking ornamental façade, no doubt; but, as is sadly so often the case where follies are concerned, with very little going on behind the pleasant, stuccoed walls.[19]

As the following sections go on to make clear, for instance, the Bill in its current form will create a regulator, and the regulator will issue codes of practice (section 2, below), and the codes will allow service providers to generate "alternative measures" to those set out in the codes, and those alternative steps will enable service providers to generate their own, proprietary and largely inaccessible safeguards (section 3), and those safeguards will be monitored by weak, ineffective regulatory measures like "information notices", "skilled persons" (section 4) and "transparency reports" (section 5) that are almost entirely opaque to anyone wishing to understand whether service providers have *actually* been doing anything to safeguard free-speech (section 6), and this fine-looking façade will stand there, proud, tall, erect… and entirely hollow, safeguarding free speech as robustly as a prison CCTV system would safeguard its wardens were the monitoring and control of that system ever to be outsourced entirely to the inmates.

2. Ofcom's codes of practice

The Government wants us to believe Ofcom is capable of "driving improvements to the safety of online service users".[20] As Ofcom has already made clear, though, it "won't be responsible for regulating or moderating individual pieces of online content".[21] That task – a rather crucial task for a regulator, really – will be left entirely to in-scope service providers to manage for themselves.

That's not to say Ofcom won't have any power at all. In lieu of direct content moderation, for instance, Ofcom will issue statutory codes of practice detailing the steps companies will need to take to fulfil what the draft Bill describes as their "duty of care" to online service users.[22] Having acquainted themselves with these codes, companies will then have the rather dubious sounding pleasure of developing "risk assessments"; that is, working out the hypothetical risk of harm posed to a hypothetical individual who doesn't actually exist but who might, hypothetically speaking, use their services, before then putting in place "appropriate systems and processes" to improve that – hypothetical, of course – user's safety.[23]

The Government claims that these codes won't just protect users from that most nebulous of toxins, online harm, but also "strengthen people's rights to express themselves freely online".[24] That's because all in-scope companies "will be required to consider users' rights, including freedom of expression online, both as part of their risk assessments and when they make decisions on what safety systems and processes to put in place on their services". In addition, they'll also need to "consider and put in place safeguards for freedom of expression when fulfilling their duties [i.e., the 'duties' set out in the codes of practice]".[25] The idea, presumably, is that because these provisions are to be applied to all in-scope companies, Ofcom's codes will "ensure transparent and consistent application of companies' terms and conditions relating to harmful content", and, as a consequence, "empower adult users to keep themselves safe online and protect freedom of expression by preventing companies from arbitrarily removing content".[26]

Given that Ofcom hasn't yet written any guidance regarding the "safeguards"

companies will need to "put in place", it is obviously unknown how robustly they're going to be looking to defend freedom of expression. A recent DCMS/Home Office press release does, however, offer a glimpse into Ofcom's thinking with the following hypothetical:

> These safeguards will be set out by Ofcom in codes of practice but, for example, might include having human moderators take decisions in complex cases where context is important.[27]

Hmm. It's not much to get excited about, is it, particularly when the protocols human moderators will be using to make those 'decisions' will inevitably be written by the service providers themselves rather than, say, a panel of independent experts working under the auspices of Ofcom. That might seem rather a trivial point, but consider this: in a recent interview, Twitter's new CEO, Parag Agrawal, cheerily admitted that he felt "[Twitter's] role is not to be bound by the First Amendment, but… to serve a healthy public conversation and our moves are reflective of things that we believe lead to a healthier public conversation". That's why, he continued, "the kinds of things that we do about this [will be to] focus less on thinking about free speech but thinking about how the times have changed".[28] Would you trust a guy like that to train moderators to "take decisions" in "complex cases" where free speech is at stake and "context is important"? The days when Twitter described itself as "the free speech wing of the free speech party" are long gone. Agrawal's not some kind of cognitive outlier, either. Silicon Valley is full of people who think just like him.[30]

What makes this all the more worrying is the fact that online providers like Twitter will risk fines and other sanctions from Ofcom if they *don't* remove material, but will be able to avoid sanctions for acting precipitously provided they meet the Bill's rather weak and watery duty to "have regard" for free speech. One obviously dislikes bandying vulgar, colloquial phrases like 'weak and watery' whilst referring to formal, governmental proposals for new legislation, but only a person wilfully blind to the realities of contemporary power could deny that there's a certain homeopathic *je ne sais quoi* about a duty that requires service providers to "have regard" for free speech.

Consider, for instance, Part 3, Chapter 2 of the most up-to-date iteration of the Online Safety Bill ("Providers of user-to-user services"). It contains 15 sections, the vast majority of which drill down into what is, at times, excruciatingly granular detail regarding the ways in which in-scope service

providers will have to demonstrate a "duty of care" in relation to their users' rights to protection from harm. Later, in Part 12, Schedule 12, the Bill also makes clear that Ofcom will have the power to fine companies up to £18 million, or 10% of global annual turnover (whichever is the higher), if they fail adequately to perform this duty.

So far, so detailed, well-defined and uncompromising.

Yet things appear quite otherwise when we turn to consider how the legislation describes the duty of care service providers will need to demonstrate in relation to everyday users' rights to online free speech.[31]

It's not that Ofcom won't be able to take enforcement action or issue hefty fines in cases where service providers are found to have failed to show regard for freedom of speech – they will.[32] It's just that the Bill has ended up defining what compliance looks like in such a loose, vaguely worded sort of a way as to ensure that almost any censorial regulatory action perpetrated by a service provider will be susceptible to retrospective discursive, stylistic and rhetorical finessing of a kind capable of allowing it to appear in any subsequent regulatory report as if it had in fact been entirely compliant with that service provider's duty to "have regard to" free speech all along.

Perhaps we can forgive the Government for the hazy, indefinite suggestion it made very early on in the legislative process, that:

> Companies will be required to *consider* users' rights, including freedom of expression online, when they make decisions on what safety systems and processes to put in place on their services [emphasis added].[33]

What's remarkable, however, is that as we move further downstream, inching ever closer to that final, fateful moment when the Bill passes into law, similarly vague, similarly abstract definitions of this 'duty of care' are still to be found at work. Chapter 2, Clause 19 of the latest iteration of the Online Safety Bill, for instance, suggests that "when deciding on, and implementing, safety measures and policies", service providers have:

> A duty to have regard to the importance of protecting users' rights to freedom of expression within the law.[34]

"Have regard to…". What will the big-tech employees over in Silicon Valley need to do in order to convince Ofcom that they've been 'having regard to', their respective users' rights to freedom of expression, do you think? Tick a box? Take the knee? Check their white privilege? Lie back and think of the Mau Mau Uprising? Attend a seminar series run by Robin DiAngelo? Insofar as Ofcom will occasionally have to judge whether online service providers have got this balancing act right, the danger is that the Bill's understanding of what a free speech duty constitutes is so weak, so vague, so poorly defined, that it could plausibly end up meaning almost anything an in-scope company wanted it to mean.

There's another problem here too. Ofcom's as-yet-unwritten codes of practice are supposed to establish blueprints from which a robust, well-regulated superstructure of compliance can subsequently be constructed. That much we know. However, large parts of that superstructure's necessary architecture have already been built and are now happily being managed and maintained by the service providers themselves. As a result, the complex, technical, algorithmic details of the proprietary content moderation systems deployed by service providers to protect users from harm, will be hard for external agencies to access, let alone understand. This, in turn, will impact their ability to understand how freedom of speech safeguards operate.

So the first question here, really, is not whether Ofcom's supposedly binding blueprints for safeguarding freedom of speech will find expression in that architecture, but whether Ofcom will ever be allowed enough access to that architecture to find out. How, for instance, will Ofcom know where, when, in what ways – *or even if* – those aforementioned "complex cases" of censorship, in which "context is important" and issues of free speech are at stake, end up reaching human moderators?[35] And what about instances where Artificial Intelligence (henceforth AI) based online content moderation systems proactively take down content – how will external auditors receive credible assurances that, during that automated process, users' rights to freedom of expression were given *meaningful* "consideration"?[36]

A little earlier we noted that the Government was keen to ensure "the effective implementation of this regulatory regime" by granting Ofcom robust enforcement tools to tackle non-compliance.[37] But if non-compliance ends up happening inside a black-boxed AI-based content moderation system, the inner workings of which Ofcom doesn't understand, can't see

and isn’t able to access, then it won’t matter how robust its enforcement tools are, will it?

3. Alternative steps, alternative measures

Still, none of that really matters, does it? Because Ofcom will be in a position to issue clearly set out, tightly defined and precisely worded codes of practice, won't it? And those clearly set out, tightly defined and precisely worded codes of practice will ensure all privately managed policing of online communication is undertaken in a transparent, standardised and auditable manner such that if and when peoples' rights to online free speech and freedom of expression are infringed, Ofcom will quickly come to know about it… *won't they?*

Maybe. And yet...

In the previous section it became clear that although Ofcom's as-yet-unwritten codes of practice will be legally binding, they will come into being in and as part of an ecosystem populated by proprietary online platforms, the technical 'back-ends' of which are closed-off from democratic scrutiny and increasingly driven by state-of-the-art, notoriously opaque and black-boxed algorithmic processes. That alone would have been bad enough, impacting Ofcom's ability to successfully perform its regulatory duties. But could things get even worse? The Bill offers up the possibility for decisions regarding the final structure of any given company's "duty of care" to be outsourced to the company in question. Ofcom will obviously set out the steps that in-scope companies need to put in place to uphold their regulatory responsibilities,[38] but, as the Government goes on to make clear, those same companies:

> May take alternative steps to those set out in the codes of practice, provided they can demonstrate to Ofcom that the steps in question are as effective as, or exceed, the standards set out in the codes.[39]

To be sure, the phrase "as effective as, or exceed, the standards set out in the codes" makes clear that "alternative steps" isn't some kind of innocuous-sounding regulatory euphemism for 'anything goes'. It's also fair to say that the Bill will create massive incentive for providers to follow

Ofcom's codes of practice – after all, strict, no-messing-about adherence will automatically ensure compliance with the required duty of care. So it's not going to be easy for service providers to tack their own, self-coded little shanty shacks to the side of this Bill's grand, legal architecture without then having to suffer the indignity of a lanyard-wearing, clipboard wielding bureaucrat from Ofcom popping by to tell them they've breached planning permission and that the whole structure will need to be demolished by midnight Tuesday week.

And yet, for all that, the language of "alternative steps" *has* been included in the Bill, and it does appear to open a little wriggle-room for service providers regarding the methods they might wish to deploy in and as part of their attempts to successfully comply with Ofcom's codes. Might that wriggle-room allow such companies to remove 'human minds' from all, or at least certain parts, of the regulatory loop, replacing 'them' with AI-based content moderation systems? And might those AI-based content moderation system subsequently come to be presented as if 'essential' to the roll-out of "alternative steps" that would be "as effective as, or even exceed, the standards set out in [Ofcom's] codes"?

Certainly, the mindset in Silicon Valley seems to be that content moderation is not so much a *socio-political* problem to be debated, argued about and resolved by people and politicians, as a *techno-scientific* problem to be assessed, fixed and patched by software engineers. That's why most of the major platforms now talk up the idea of AI-based software that will one day be capable of identifying problematic content more quickly and more fairly than human reviewers – during his Congressional testimony in 2018, for instance, Facebook CEO Mark Zuckerberg referred again and again to AI, glossing the many technologies groups under this generic heading as the future solution to Facebook's political problems. But of course, the fact that these systems are becoming increasingly automated has posed, and will continue to pose, significant challenges from an auditing perspective, not least because they lack public transparency and meaningful democratic accountability.[40] As a recent research paper noted:

> It will become significantly more difficult to decipher the dynamics of takedowns (and potential human rights harms) around some policy issues when the initial flagging decisions were made by automated systems, and the specific criteria by which those initial decisions were made remain unknown. From a user perspective, there is little transparency around whether (or to what extent) an

> automated decision factored into a takedown. The specific functionalities of these systems are left intentionally vague, and the databases of prohibited content remain closed off to all – including, worryingly, trusted third-party auditors and vetted researchers.[41]

If service providers' users are granted such little transparency regarding automated takedowns, one wonders what level of transparency Ofcom might be afforded if ever it fancied assessing whether a part or fully automated content moderation system that featured within a service provider's "alternative steps" really was "as effective as, or exceed[ed], the standards set out in the codes". One rather suspects it would be forced to rely less on knowledge of what proprietary, black-boxed AI systems were *actually* doing, than on knowledge of what the company that owned those black-boxed AI systems told them that they were doing. The two are not, of course, necessarily the same thing.

So does the language of "alternative steps" offer up a way for companies to deploy algorithmic systems that could then be positioned rhetorically as having technically achieved, *or perhaps even exceeded*, the standards set out within Ofcom's codes of practice? It would be concerning if this were the case – in that no independent auditors or research teams could possibly verify such claims, Ofcom would be placed at a huge disadvantage vis-à-vis the entities it would supposedly be regulating.

But *will* this ever be the case? Will the language of "alternative steps" become something of a get-out-of-jail-free card for the masters of the universe? Might the largest in-scope service providers end up crafting their own AI-based "alternative steps"? And would any of us be particularly surprised if those steps ended up fitting rather more snugly around each respective company's specific financial and political objectives than around Ofcom's regulatory regime?

We obviously can't know with any certainty what *practical* significance this language will have once the Bill passes into law. All we can say for now is that if – and it is still a big "if" – companies *do* end up taking "alternative steps", and those "alternative steps" *do* end up including black-boxed proprietary AI-systems, it will be difficult, if not impossible, for Ofcom to ensure consistent application of its own, original codes of practice. Indeed, under such a scenario, Ofcom wouldn't so much be *policing* the adequacy of a company's free speech safeguards as *trusting* people like Parag Agrawal to do the right thing by free speech whilst

curating those “healthier public conversations” they all seem so keen on. You know the type of “healthier public conversations” I’m talking about, don’t you?

…

That’s right, those ones; the ones in which Agrawal says he wants Twitter to, er, “focus less on thinking about free speech but thinking about how the times have changed”.[42]

By the way, the big-picture problem here isn’t difficult to discern, is it? *Gross technical disparities*. Where service providers are allowed to mould proprietary systems around “alternative steps” to those set out in Ofcom’s codes, those disparities will likely be perpetuated. Existing, longstanding proprietary algorithmic systems will simply be rubberstamped as they continue to chug away out of sight of any meaningful regulatory process.

There were obviously always going to be technical disparities between Ofcom and the largest online service providers. But the Bill could and should have done more to explore innovative and creative ways to minimise them, even if only for the purposes of external audit and inspection. It’s true, of course, that the latest iteration of the Bill requires online providers to update their own self-produced risk assessments as and when they make updates to their systems. And yet, as the label “risk assessment” suggests, documents of that kind will inevitably be focused more on the possible downstream, existential, user-focused *effects* of systemic change, rather than their upstream, technical, algorithm-focused processed *causes*. To get at the latter, one would need *meaningful*, provided-for-by-legislation opportunities for expert external oversight of any such proposed updates and alterations to systems.

It isn’t like we haven’t been here before, either. When it comes to an exercise like tax collection, for instance, HMRC regularly bends the rules to accommodate those rather eccentric interpretations of UK tax law so beloved of large, footloose, and immensely wealthy global companies. That’s partly for *financial* reasons, of course: the companies in question are too important to the UK’s economy for HMRC to annoy them. But it’s also for *technical* reasons: those companies employ the best accountants money can buy, which means HMRC is rarely able to lay a glove on them. A so-called “sweetheart deal” struck with the American multinational Goldman Sachs in 2010, for instance, was described at the time as

'rewarding' the bank for several years of failing to pay tax. Yet the deal was signed off by HMRC. Why? Because according to the UK Government's then permanent secretary for tax, Dave Hartnett, Goldman Sachs "went off the deep end" at the suggestion they should pay what they owe, and subsequently threatened to withdraw from the Government's recently introduced banking code of practice.[43] Again, more recently, HMRC was forced to settle a dispute with the multinational conglomerate GE over a $1.1 billion tax scheme that the company ran from 2004-15. The final bill payable by GE, an employer of approximately 9,000 people around the UK? Just $112 million, or 10% of the potential tax savings the company generated during the period in question.[44]

We could go on. And on. But it's not difficult to spot that and see how a similar dynamic might well emerge between Ofcom and the largest, riskiest category 1 service providers as and when they decide to pursue their own "alternative steps".

4. Information notices and "the skilled person"

To a certain extent, the Bill recognises the dangers presented by these technical disparities. It makes clear, for instance, that Ofcom will need "more expertise in technology, especially emerging tech and the use of Artificial Intelligence and the ways both drive commercial and consumer change".[45] It also confirms that, "the Government will continue to explore a range of measures to support the rapid development of the safety tech market", and that one such measure will be the delivery of "a new £2.6m project to prototype how better use of data around online harms can lead to improved Artificial Intelligence systems and deliver better outcomes for citizens".[46] £2.6 million. It's what you might call a warm-hearted but essentially futile gesture, isn't it? The political equivalent of garage-bought flowers for a wife whose bags have long since been packed and are now tucked away in the boot of a waiting taxi. To put the Government's proposed level of spending into context, Facebook recently announced plans for $29-34 billion in capital expenditure on AI, servers and data centres in 2022. (The year before that, by the way, it spent $19 billion.)[47]

In truth, Ofcom's executive powers would need to be increased, and increased drastically, for it to stand any chance at all of holding online providers to account. As a minimum, it would need access to data and code and the authority to investigate algorithms as possible drivers of unwarranted censorship.[48] It would also need the power to query systems engineers and product managers to determine the relative impact of design and features choices on a company's ability to safeguard free speech.[49] It's not that there's any language in the Bill that specifically excludes Ofcom from initiating audits of service provider algorithms; it's just that recent commentary based on the draft Bill suggests Ofcom doesn't view algorithmic audit as a power in its toolkit.[50] The language in Part 7, Chapter 4, clauses 85-102 of the Bill does in fact state that Ofcom can serve an "information notice" to a service, requiring a person to provide information which Ofcom believes that person has or is able to generate or obtain.[51] In addition, Ofcom is to be granted powers to "appoint a skilled person" to help it with an investigation.[52] It sounds a bit thin, doesn't it? Could one person really

make *that* much of a difference?

Information notices, extra funding, skilled persons: nice enough if you're fond of ticking a few boxes every now and then. But hardly likely to assist Ofcom in *actually* assessing whether the *actual* safeguards companies have put in place to protect freedom of speech will *actually* be upheld in any meaningful way. Which does rather beg the question of what, if anything, Ofcom has in its regulatory locker that's likely to prove capable of holding online providers to account for the adequacy of their free speech safeguards.

5. The transparent, self-performed audit

Consider a scenario in which Ofcom wanted to assess the adequacy of the free-speech safeguards in place at a category 1 service provider like Facebook. Let us suppose it had been prompted to take this action by user complaints data indicating that the company's proactive, AI-based content moderation systems were censoring information that didn't meet some threshold for consideration as 'harmful' or 'misinformation'.

I know what you're thinking, of course. And you're right: it is a laughably implausible scenario. The sight of a regulatory watchdog perched atop a war chest known to contain as much as £2.6 million, barking out demands that companies build their own AI-based online moderation systems before then managing them in accordance with their own "alternative steps" would alone surely suffice to check the censorial ambitions of any wayward, multi-billion-dollar bigtech giants. And when we add that this digital Cerberus is known to include within its vast arsenal weapons such as "the skilled person" and "the information notice" we think we have said everything as to the folly of the Mark Zuckerbergs of this world coming on strong with a spot of the old semiotic argy bargy. Ofcom would crush the man as you or I would crush an insignificant worm.

Just for the purposes of argument, though, let us suppose that this laughably implausible scenario had indeed come to pass; that Ofcom felt it had no choice but to start gathering information regarding the free speech safeguards in place at a service provider like Facebook. How would it set about doing so? Information notices and skilled persons aside, the Government apparently believes that Ofcom will be able to hold companies like Facebook to account on the basis of what they're calling "transparency reports".[53] Not that these audits will be undertaken by external parties, you understand. Oh no. That would be dangerous. The researchers might be biased. Spies. Working for rival companies. Politically motivated. With axes to grind. That's why, in a breath-taking display of intellectual leadership, the Government has decided that the only reasonable, plausible, logical and robust way to ensure online companies implement effective freedom

of expression safeguards will be to require those same online companies to produce transparency reports on… well, on themselves, really.[54]

In practice, companies will be required to submit their transparency reports to Ofcom on an annual basis. Specifically in relation to safeguarding, they will be required to "conduct and publish up-to-date assessments of their impact on freedom of expression and demonstrate they have taken steps to mitigate any adverse effects".[55]

Will they work in Ofcom's favour?

It's fair to say that transparency reports are a tried, if not altogether trusted, regulatory mechanism. 'Tried' in that toothless online regulators have in the past tended to fall back on voluntarily prepared transparency reports simply for want of the power to beg the companies they were at least nominally supposed to be regulating for anything else.[56] 'Not altogether trusted', however, in that they rely on companies performing actions they very rarely want to perform: opening themselves up to external scrutiny, sharing data, sharing the metrics via which they measure the effectiveness of their automated and human content moderation systems, sharing information on their policies with regulators and revealing certain aspects of their proprietary internal systems and processes to outside agencies.[57] Even in cases where companies have previously coughed up transparency reports, they've shown a worrying tendency to include certain types of information (particularly in relation to what kinds of content they remove from their platforms) that researchers can't externally validate. Nor have they been averse to a spot of misdirection from time to time, either, feeding researchers and regulators self-selected statistics that couldn't transparently answer the questions being asked of them.[58]

In the parlance of the criminal justice system, then, we might most tactfully describe the voluntarily prepared transparency report as having a 'background'. Not that that necessarily matters when it comes to the draft Online Safety Bill. After all, the legal ecosystem proposed therein will *require* online companies to prepare these annual audits. But then, so what? In the absence of any prescriptive regulatory guidelines, companies will be free to prepare those reports in any way they choose. All the draft Bill requires companies to compile are *transparency-reports*. What we don't yet know is whether the specifics of Ofcom's codes will require the compilation of *reports-that-are-transparent*.

The Government does of course claim that transparency reports will "remove the risk that online companies adopt restrictive measures or over-remove content in their efforts to meet their new online safety duties".[59] But that claim only makes sense if you believe you're dealing with companies that will compile transparency reports stocked full of data capable of allowing external experts, researchers and regulators to generate transparency regarding the appropriateness or otherwise of their actions in relation to the safeguarding of freedom of speech.

Is the Government right to believe that? It all depends on the level of detail Ofcom's as-yet-unwritten codes will require companies to provide in their transparency reports. Will they produce precisely worded, technically specific, heavily prescriptive guidelines that channel companies along well-defined and unavoidable lines? Or will they instead formulate looser, indicative clusters of possible 'areas' or 'themes' that service providers 'should consider' including, thus leaving them with a good deal of wriggle-room when it comes to the details they include?

The only clue we've had so far as to how Ofcom might approach this issue appears in part four of the full Government response to the consultation period. Therein, we find "an indicative list" of the high-level categories of information companies "might" need to include in their transparency reports.[60] It's rather an interesting list. If we were looking to understand the types of information Ofcom "might" require online companies to provide regarding the steps they've taken to protect users from online harm we'd be sitting pretty: eight of the list's nine bullet points cover that topic. But when it comes to the type of information companies "might" be required to provide in relation to their free speech safeguards, there just doesn't seem to be the same level of Governmental concern, with one bullet-point covering that topic. What might that disparity tell us, do you think, about the Government's overarching priority within the Online Safety Bill?

Still, this one, solitary bullet point does at least give us an 'indicative' idea that in-scope companies might be required to provide information regarding "the measures and safeguards [they have in place] to uphold and protect fundamental rights; to ensure decisions to remove content, block and/or delete accounts are well founded, especially when automated tools are used, and that users have an effective route of appeal". But if that's the type of *content* Ofcom might wish to collect, then it's important to understand the form that that information is likely to take in and as part of a transparency report.

6. The rhetoric of numbers

There are, broadly speaking, two forms in which data may be presented: qualitatively (with words and pictures) or quantitatively (with numbers and graphs). Service providers have in the past shown a predilection for data of the latter type, producing reports populated almost entirely with metrics like absolute values (e.g. the number of takedowns per quarter and/or the amount of content removed by a platform) and proportional metrics (e.g. the percentage of all content which violates a service's policy).[61] 'Almost entirely' in the sense that they've tended to focus on summary statistics at the expense of the type of qualitative, contextual information that would have allowed researchers to 'situate' the data as part of some overall interaction.[62]

It's difficult to see anything changing in that regard. Putting that another way, the Online Safety Bill's proposed requirement for service providers to report on the adequacy of their free speech safeguards won't have any of them rushing to provide qualitative data. Not that there's anything particular controversial about that. Service providers will be required to report complex data to regulators, and numbers have the power to reduce complexity. And yet…

And yet any such reduction of complexity can be neither theoretically nor ideologically innocent.[63] Raw data can, after all, be collected, combined, processed, analysed in myriad ways. That it is not; that it is only ever collected, combined, processed, analysed in certain extremely specific ways by certain people working for certain organisations as part of certain projects, suggests that there is in fact a *rhetoric* to numbers. And it is this rhetoric that service providers will undoubtedly look to exploit whilst preparing their transparency reports.

Consider the fact that a transparency report "might" be required to include information regarding the "measures and safeguards" a company has in place "to ensure decisions to remove content, block and/or delete accounts are well founded". It would of course be perfectly possible to meet this requirement via the provision of qualitative information regarding, say, the blocking or removal of content labelled as 'misinformation'. The problem

is that any such effort would run to rather a few pages. That's the problem with qualitative data. By its very nature it's particularistic. It looks to capture the specific contexts in which individual instances of a phenomenon – in this case 'misinformation' – unfold, and, in addition, the particulars of who said what, how it was said, when it was said, in what order interlocutors spoke, to whom something was said, where an interaction took place and so on. It's not necessarily that qualitative data *wouldn't* refer to a global category like 'misinformation'; it's just that it would do so whilst discussing individual instances of action that were potentially capturable under that category.

This tendency of qualitative data towards the particular, the specific and the local would be all the more pronounced in the case of a complex, contested and controversial category phenomenon like 'misinformation'. Even if we *were* to accept that such a thing existed, instances of 'it' would be incredibly hard to capture. It might emerge within many structurally different forms of online interaction (synchronous, asynchronous, or quasi-asynchronous)[64] within many different modalities of online communication (visual, video-based, textual)[65] and across many different communicational contexts (one-to-one chats, multi-party settings, or massive, open group settings).[66] From a qualitative perspective, then, to report on the "measures and safeguards" a company has in place to ensure decisions to remove content are well founded is to foreground real-world data drawn directly from that online provider's platform, and to do so such that it becomes possible to see 'it' as part of the wider, interactional context in which it unfolded.

There would, in other words, be no discussion of 'misinformation' without discussion of, say, a fired-up, politically engaged Sharon from Doncaster whose content often gets taken down by Twitter for violating their policies around 'misinformation'. Real-world data of that kind would capture an entire sequence of interaction, in this case from Sharon's initial Tweets and interactions right through to Twitter's eventual removal of her Tweets. It's in that way that a qualitative transparency report would seek to demonstrate whether, if, when, how and for what reasons the "measures and safeguards" in place were "well-founded".

As we've already noted, however, that sort of approach doesn't cover the regulatory ground particularly quickly. And if it's speed of coverage you're after, then quantitative data will always be your go-to presentational form. Whereas qualitative approaches interest themselves in the complexities

of individual cases [*Sharon from Doncaster did this, someone reacted to Sharon this way, Sharon responded like this… etc.*], quantitative approaches sweep people up into groups [*i.e., the group "Purveyors of Misinformation"*] alongside other apparently similar people [*Sharon from Doncaster and Donna from Grimsby and Barry from Northampton and… and… and…*] before deleting every particular, specific contextual detail about those people [*names, biographies, characteristics, specifics of the argument, specifics of the interactional situation*] and leaving us not with 'people' but 'cases'; that is, cases that can be added, subtracted, divided, multiplied, counted (etc.) to create absolute values or proportional metrics [*4,563 posts this month were found to have been produced by Purveyors of Misinformation; 96.4% of those posts were automatically taken down; 3.4% were… and so on*].

That's why quantitative approaches to transparency reporting make it easy to provide information about things like "measures and safeguards". A potentially infinite number of cases of content blocking or removal become reportable within just a few metrics spread out across just a few short paragraphs. One quite straightforward way to cover an entire months' worth of data regarding content blocking might even look something like this:

> Amount of content per month violating a company's policy around 'misinformation': 14,530 posts [*Absolute Value]*
> Percentage of platform content violating a company's policy on 'misinformation' that is subsequently subject to automated take-down: 3.7% [*Proportional metric*]

You will note, of course, that in far less textual space than it took to report the activities of just one 'person' – i.e., the events leading up to the takedown of Sharon's Tweets – we have managed to report a total of 14,350 'cases', and, in fact, had more than enough space left over to decompose that absolute value into a separate proportional metric.

To quantify, though, is to engage in processes of simplification that are capable of serving purposes quite other than literal truth telling. In the case of the transparency reports online companies will be required to prepare, for instance, we might say that quantification has the potential to *constitute* and *depoliticise* a phenomenon like 'misinformation'. Let us consider each process in turn.

One of the hypothetical proportional metrics we listed earlier was 3.7%; that is, "the percentage of platform content violating a company's policy around 'misinformation' that is subsequently subject to automated take-down". It's a statement that immediately brings to mind the philosopher John Austin's work on "performatives". As he pointed out in his wonderfully anarchic *How to do Things with Words* (1962), performatives were rather an unusual form of speech-act. They didn't just describe a state of affairs but also, and at the same time, brought that state of affairs into being: e.g. the statement "I pronounce you man and wife", when uttered by a vicar whilst standing in front of a bride and groom in a church.[67] In the current socio-political climate, a statement like "3.7% of all content this month violated our service's policy around 'misinformation'" would undoubtedly have a similar performative force were it to be uttered by a service provider as part of a transparency report submitted to Ofcom. Therein, the percentage figure, 3.7%, would not so much inscribe the pre-existing reality of 'misinformation' as *constitute* it.[68]

A controversial object, 'misinformation', an object many people don't actually believe exists, would in this way be rendered as an objective, definitive, statistically measurable 'thing'. Political arguments as to whether that which the Government has defined so clumsily as 'misinformation' even exists, about whether anything that meets that definition should ever be subject to online takedown policies *of any kind,* would subtly be displaced into an instrumental argument as to whether the figure of 3.7% was accurate or inaccurate; whether the company's metrics were adequate or inadequate; whether the… and so on.

Over time another form of displacement would occur too. Methodological arguments regarding how best to capture 'it' [*because it would, of course, have become an 'it' by this stage*] would end up becoming a substitute for democratic discussion regarding the principles of free speech and freedom of expression: questions as to whether audits, and, by implication, metrics, were the right way to engage with the many, complex issues surrounding both free- and hate-speech in online environments would be drowned out because, well… 'Look! There's 3.7% of this thing and we need to do something about it!' It is in this way that the 'fact' of 'misinformation' would be constituted; its very existence captured in and through the statistical indicator that purports to measure 'it'.

But metrics have the power to *depoliticise* too.[69] They allow for many, potentially disparate instances of 'misinformation' to be brought into relation

with one another. To record those disparate instances from a qualitative perspective would of course be to record the specificities of each particular instance of misinformation encountered [*… and Sharon from Doncaster, and Donna from Grimsby, and Barry from Northampton, and… and… and…*], whilst at the same time acknowledging that each person had engaged in a particular form of online interaction across a particular communicational context. In that sense, qualitative data reports would be 'mutable mobiles'; that is, texts capable of transportation [*they are mobile*], but only where the meaning of the data contained therein would be subject to multiple, competing, sometimes even commensurate, interpretations as it got moved across contexts, communities of practice, cultures, social classes, political divides and so on [*they are mutable*]. One person's evidence of 'misinformation' would only ever be another person's 'free speech', 'hate speech', 'legitimate debate' or 'counter-speech'.

Specificities, details, complexities… none of that matters when you enter the world of metrics. Sharon and Donna and Barry are 'people' no more. They're 'cases' caught amongst many other cases captured by the percentage figure of 3.7%. It's in this sense that quantitative transparency reports may be considered "*im*mutable mobiles";[70] that is, texts that can be transported anywhere in the world [*they are mobile*] without the data they contain ever-changing shape or meaning [*they are immutable*]: 3.7% is 3.7% is 3.7% whoever you are, whenever you are and wherever you are.[71]

On the one hand, then, qualitative data reports open a definition like 'misinformation' to discussion and debate, affording its readers an opportunity to inspect the processes that led to a decision to classify any given interaction as 'misinformation'. It's impossible for a qualitative report to present some once-and-for-all definition of 'misinformation' precisely because the details it contains can be interpreted in multiple ways by multiple readers. At best, such a report would contain an *assertion* that certain episodes of interactional data *should* be treated as 'misinformation'… but nothing more. Quantitative data, on the other hand, shuts down those discussions, concealing the decisions a service provider has made regarding what was and what was not 'misinformation' in and part of the constitution of a statistic like '3.7%'. It hides the complex array of contestable, controversial, political judgements and decisions that went into its production.[72] A quantitative transparency report tells us only that there is 3.7% of it, and subsequent debates can only proceed from the fact that 'it' exists. This is to depoliticise what is, *or at least should be*, a whole area of political judgement. It is in this way that

statistics, and the specialist knowledge and professional techniques of the service providers associated with their production, come to be implicated in the creation of a domain where technical expertise dominates political debate.[73]

Ultimately, then, the rhetorical force of the metric is that it constitutes and de-politicises; that is, it distils simple, uncomplicated categories like 'misinformation' from all manner of complex, contextually specific interactions [*… and Sharon from Doncaster, and Donna from Grimsby, and Barry from Northampton, and … and … and …*] before erasing all evidence that such a process took place [*…99.6% of posts were misinformation…*]. The metric is constituted: it becomes a 'fact' [*… 99.6% of this, 0.4% of that, 6.7% of the other…*]. The fact is subsequently de-politicised: it is freed from any suggestion of human bias, of human intervention [*…3.7%, follow the science, the data speaks for itself, facts are facts…*]. The only problem – and it is rather a big problem – is that this process deletes exactly the type of qualitative data Ofcom would require in order to assess whether speech was being over-blocked, unnecessarily censored or over-zealously shadow-banned by service providers.

Flitting back and forth between qualitative and quantitative data in the way we have during this section is to trace the contours of digital power in the twenty-first century. Where qualitative data is capable of exposing the machinations of big-tech giants as they look to jettison free speech in favour of "healthy public conversations", quantitative data simply, effortlessly conceals them; where qualitative data is capable of opening the process of defining nebulous, politically controversial, and entirely contestable phenomenon like 'misinformation' to democratic argumentation, quantitative data simply, effortlessly closes that possibility off. It is for these reasons that transparency reports will almost certainly turn out to be *quantitative* reports; and it is for these reasons, too, that they will provide Ofcom with a form of transparency that is almost entirely opaque to anyone wishing to know how, when, whether, where – *or even 'if'* – service providers are doing anything at all to safeguard online free speech.

References

[1] Department for Digital, Culture, Media & Sport (2022). *Online Safety Bill*. London. HMSO. (58/2). Available at: Online Safety Bill publications – Parliamentary Bills – UK Parliament [accessed 29 March 2022].

[2] Department for Digital, Culture, Media & Sport (2022). Press Release: World-first online safety laws introduced in Parliament. 17 March. Available at: World-first online safety laws introduced in Parliament – GOV.UK (www.gov.uk) [accessed 28 March 2022].

[3] The phrases "priority content that is harmful to children" (clause 53) and "priority content that is harmful to adults" (clause 54) appear in the latest iteration of the Bill. Both are effectively granted the same legal definition, as "content of a description designated in regulations made by the Secretary of State as priority content" (for children, see clause 53, sub-section 3; for adults, see clause 54, sub-section 2).

Effectively, then, the only #legal but harmful' content the latest iteration of the Bill requires social media companies to remove will be that caught by "regulations made by the Secretary of State"; that is, via a Statutory Instrument (i.e., secondary legislation). On the face of it, that means the Government has now given up on the idea of outsourcing the prohibition of 'legal but harmful' content to service providers… which is obviously a good thing. What's not such a good thing, however, is that as clause 53, sub-sections 3 and clause 54, sub-section 2 both make clear, the Bill will instead look to create a mechanism via which the Secretary of State – whosoever that may be, now or in the future – will be able to identify 'legal but harmful' content in and via that aforementioned Statutory Instrument.

Secretaries of State obviously won't have carte blanche when it comes to identifying such content, since they won't be able to compel social media companies to disregard the protections for "content of democratic importance" and "journalistic content" included in the current version of the Bill. But they will have some latitude. For instance, a future Secretary of State could potentially decree that gender critical beliefs were, in the words of the judge in the original Forstater v CGD Europe tribunal hearing, "unworthy of respect in a democratic society". If they ever did end up categorised as such, then they wouldn't be entitled to the Bill's "content of democratic importance" protection, which would obviously leave them prey to any and all social media companies wishing to remove 'that sort of thing' from their platforms.

[4] Department for Digital, Culture, Media & Sport (2022). *Press Release: World-first online safety laws introduced in Parliament*. 17 March. Available at: World-first online safety laws introduced in Parliament - GOV.UK (www.gov.uk) [accessed 28 March 2022].

[5] Department for Digital, Culture, Media & Sport (2019). *Online Harms White Paper*. CP57. London: HMSO. Available from: Online Harms White Paper - April 2019 - CP 57 (publishing.service.gov.uk) [accessed 3 March 2022].

[6] Department for Digital, Culture, Media and Sport. (2020) *Online Harms White Paper: Full government response to the consultation*. CM354. London. HMSO, para 4.43. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[7] Department for Digital, Culture, Media & Sport (2021). *Draft Online Safety Bill*. CP405. London: HMSO. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf [accessed 3 March 2022].

[8] Ofcom (2021). *Ofcom to gain new online safety powers as Government bill published today*. Available at: Ofcom to gain new online safety powers as Government bill published today - Ofcom [accessed 3 March 2022].

[9] Ofcom will in fact constitute the first overall regulator with responsibility for online content in the UK. See House of Lords Select Committee on Communications (2019). *Regulating in a digital world*. 2nd Report of Session 2017-19. HL Paper 299. London. HMSO. Available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldcomuni/299/299.pdf [accessed 3 March 2022].

[10] Joint Committee on the Draft Online Safety Bill (2021). *Corrected oral evidence: consideration of government's draft Online Safety Bill, Monday 1 November 2021, Evidence Session No. 16, 2:30 pm*. Available at: https://committees.parliament.uk/oralevidence/2934/pdf/ [accessed 3 March 2022]; for context, also see Evening Standard (2021). *Online Safety Bill provides Ofcom with 'right powers to keep big tech in check'*. Available at: Online Safety Bill provides Ofcom with 'right powers to keep big tech in check' | Evening Standard [accessed 3 March 2022].

[11] HM Government (2020). *Online Harms White Paper – Initial consultation response*. London: HMSO, paras. 11 & 24. Available at: http://data.parliament.uk/DepositedPapers/Files/DEP2020-0111/Online_Harms_White_Paper-Initial_consultation_response.pdf [accessed 3 March 2022].

[12] Tettenborn, A. (2020). Ofcom is a menace to our freedom of speech. *The Critic*. 24 September 2020. Available at: Ofcom is a menace to our freedom of speech | Andrew Tettenborn | The Critic Magazine [accessed 3 March 2022].

[13] On the algorithmic power of the so-called "masters of the universe", see Gorwa, R., Binns, R., Katzenbach, C. (2020). Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society*, Jan-June, 1-15.

[14] Jones, W. (2021). Why is Ofcom Suppressing Covid Information Based on the Advice of a Biased 'Fact-Checker' Funded by Google, Facebook and George Soros? *The Daily Sceptic*. Available at: Why is Ofcom Suppressing Covid Information Based on the Advice of a Biased 'Fact-Checker' Funded by Google, Facebook and George Soros? – The Daily Sceptic [accessed 16 April 2022]; also see Thompson, D. (2021) Stonewall's influence on BBC and Ofcom revealed. *BBC News*. Available at: https://www.bbc.co.uk/news/uk-58917227 [accessed 3 March 2022]; Aitken, R. (2021) Don't be fooled again – Ofcom is no free-speech saviour. *Daily Telegraph*. Available at: https://www.telegraph.co.uk/tv/0/dont-fooled-ofcom-no-free-speech-saviour/?li_source=LI&li_medium=liftigniter-onward-journey [accessed 3 March 2022]; Tettenborn, A. (2021) Another Ofcom power grab. *Spiked Online*. Available at: https://www.spiked-online.com/2021/07/28/another-ofcom-power-grab/ [accessed 3 March 2022]; Davenport, N. (2021) Ofcom wants to No Platform trans-sceptics. *Spiked Online*. Available at: https://www.spiked-online.com/2021/01/06/ofcom-wants-to-no-platform-trans-sceptics/ [accessed 3 March 2022].

[15] Full Fact (2022). About us – funding. *Full Fact*. Available at: Funding - Full Fact [accessed 16 April 2022].

[16] At the brutalist level of pure, amoral functionality, what links the many, apparently disparate strategies for censoring online hate and/or 'misinformation' so beloved of the largest service providers is the fact that they all apply varying levels of 'friction' to certain types of 'undesirable' information. Following the Alan Turing Institute (*Understanding Online Hate: VSP regulation and the broader context*. Available at: Report from The Alan Turing Institute - Understanding Online Hate (ofcom.org.uk) [accessed 16 April 2022]),

friction can be defined as: “The degree of resistance that content encounters in order to be published and found, seen , shared and engaged with by audiences” (p.80). At the highest level of friction, for instance, service providers can ban users from their platforms. At a much lower level of friction, service providers can impose search constraints to ensure the removal of certain types of content from any and all search results subsequently provided through that platform (pp. 82-85).

[17] On this point see The Free Speech Union (2021). *You’re on Mute: The Online Safety Bill and what the Government should do instead*. Available at: https://freespeechunion.org/youre-on-mute-the-online-safety-bill-and-what-the-government-should-do-instead/ [accessed 3 March 2021].

[18] Dame Melanie’s comments came during a Digital, Culture, Media and Sport Select Committee, and were in response to a question from the MP John Nicolson that began: “I notice that the BBC seems to be under the impression that it needs to ‘balance’ all its reports about trans issues now, by calling in *transphobic groups* like the so-called LGB Alliance to give a counter argument…” (emphasis added). Nicolson was of course presupposing that this pejorative categorisation of LGB Alliance was factual, unproblematic and universally accepted. Dame Melanie’s response suggests that she shares this same implicit assumption: “I think that is a very good point and actually *a very good example* of something that we’ve been talking to Stonewall about actually – about how can the broadcasters, when they do feel they need to bring balance into a debate, do it in an appropriate way, *rather than in the way you just described, which would be extremely inappropriate*” (emphasis added). Also see Marlborough, C. (2020). ‘Entirely inappropriate’ to quote LGB Alliance on trans issues, says Ofcom chief. *The Scotsman*. [Online] 16 December. Available from: ‘Entirely inappropriate’ to quote LGB Alliance on trans issues, says Ofcom chief | The Scotsman [accessed 28 March 2022]; also Tettenborn, A. (2020). Ofcom threatens diversity of opinion. *The Critic*. [Online] 23 December. Available at: https://thecritic.co.uk/ofcom-threatens-diversity-of-opinion/ [accessed 3 March 2022].

[19] As noted by the Joint Committee on the draft Online Safety Bill, the Draft Online Safety Bill is really rather “complex”. See *Report of session 2021-22, 14 December 2021*, HL Paper 129, HC 609, p. 21. Available at: https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/129.pdf [accessed 3 March 2022].

[20] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 4. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[21] Ofcom (2020). Ofcom to regulate harmful content online. Available at: Ofcom to regulate harmful content online - Ofcom [accessed 3 March 2022].

[22] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 2.48. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[23] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 26. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022]; also see Department for Digital, Culture, Media & Sport (2021) *Draft Online Safety Bill*. CP405. London: HMSO, Clause 29(3). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf [accessed 3 March 2022].

[24] Gov.UK (2021). *Press Release: Landmark laws to keep children safe, stop racial hate*

and protect democracy online published. Available at: https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published [accessed 3 March 2022].

[25] Gov.UK (2021). *Press Release: Landmark laws to keep children safe, stop racial hate and protect democracy online published*. Available at: https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published [accessed 3 March 2022].

[26] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para. 5.32. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[27] Gov.UK (2021). *Press Release: Landmark laws to keep children safe, stop racial hate and protect democracy online published*. Available at: https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published [accessed 3 March 2022].

[28] Bernstein, B. (2021). Incoming Twitter CEO said company should 'focus less' on free speech. *National Review*. November 29. Available at: https://www.nationalreview.com/news/incoming-twitter-ceo-said-company-should-focus-less-on-free-speech/ [accessed 3

[29] Halliday, J. (2012). Twitter's Tony Wang: "We are the free speech wing of the free speech party". *The Guardian*, 22 March. Available at: Twitter's Tony Wang: 'We are the free speech wing of the free speech party' | Changing Media Summit | The Guardian [accessed 16 March 2022].

[30] See for instance, Bokhari, A. (2019) 'Big-tech vs free speech: Breitbart exposed the masters of the university in 2018 Townhall. *Breitbart*, 5 May. Available at: Big Tech vs Free Speech: Breitbart Exposed the Masters of the Universe in 2018 Townhall [accessed 16 April 2022].

[31] The term 'everyday user' is deployed here to distinguish the heavily qualified and caveated rights to freedom of speech and expression the Bill allocates to 'normal' adults and children, from the ever-so-slightly less qualified and caveated rights it grants to those it considers to be producers of 'journalist content'. (Quite what this distinction will mean to and for the future of 'citizen journalism' is anyone's guess). See Department for Digital, Culture, Media & Sport (2022). *Online Safety Bill*. London. HMSO. (58/2), clause 16, Available at: Online Safety Bill publications - Parliamentary Bills - UK Parliament [accessed 29 March 2022].

[32] See for instance Department for Digital, Culture, Media & Sport (2022). *Online Safety Bill*. London. HMSO. (58/2), clause 111(2). Available at: Online Safety Bill publications – Parliamentary Bills – UK Parliament [accessed 29 March 2022].

[33] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para. 2.10. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[34] Department for Digital, Culture, Media & Sport (2022). *Online Safety Bill*. London. HMSO. (58/2), Chapter 2, Clause 19. Available at: Online Safety Bill publications - Parliamentary Bills - UK Parliament [accessed 29 March 2022].

[35] See Dias Olivia, T., Antonialli, D., Gomes, A. (2020) 'Fighting hate speech, silencing drag queens? Artificial Intelligence in content moderation and risks to LGBTQ voices online', *Sexuality & Culture*, 25, pp: 700-732; Binns, R., Veale, M., Kleek, M., Shadbolt, N. (2017) 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation', *International conference on social informatics*, pp. 405-415. Berlin: Springer.

[36] Llanso, E. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, Jan-June, 1-6.

[37] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 4.43. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[38] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, part 2, Box 12. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[39] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 2.48. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[40] On this point see for instance, Kaye, D. (2019). *Speech Police: The Global Struggle to Govern the Internet*. New York, NY: Columbia Global Reports; also Suzor, N.P. (2019). *Lawless: The Secret Rules that Govern our Digital Lives*. Cambridge, UK: Cambridge University Press.

[41] Gorwa, R., Binns, R., Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, Jan-June, page 11.

[42] Bernstein, B. (2021). Incoming Twitter CEO said company should 'focus less' on free speech. *National Review*. November 29. Available at: https://www.nationalreview.com/news/incoming-twitter-ceo-said-company-should-focus-less-on-free-speech/ [accessed 3 March 2022].

[43] Independent (2013). "Sweetheart" deal between HMRC and Goldman Sachs was struck to save Government embarrassment, court hears. *Independent* [online]. 'Sweetheart' deal between HMRC and Goldman Sachs was struck to save Government embarrassment, court hears | The Independent | The Independent [accessed 3 March 2022].

[44] TaxWatch (2021). TaxWatch calls for scrutiny over "sweetheart" tax deal between HMRC and GE. Available from: TaxWatch calls for scrutiny over "sweetheart" tax deal between HMRC and GE – TaxWatch (taxwatchuk.org) [accessed 3 March 2022].

[45] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 3.16. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[46] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 5.5, Box 20. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[47] Moss, S. (2021). Facebook plans huge $29-32 billion capex spending spree in 2022, will invest in AI, servers, and data centres. *Data Centre Dynamics*. Available at: https://www.datacenterdynamics.com/en/news/facebook-plans-huge-29-34-billion-capex-spending-spree-in-2022-will-invest-in-ai-servers-and-data-centers/ [accessed 3 March 2022].

[48] On this point the Free Speech Union agrees with the Ada Lovelace Institute. See Joint Committee on the Draft Online Safety Bill (2021). *Written Evidence Submitted by the Ada Lovelace Institute*. OSB0101. Available at: https://committees.parliament.uk/

writtenevidence/39256/html/ [accessed 3 March 2022].

[49] For a cogent summary of the need for this power, see Dr Amy Orben's written evidence to the Joint Committee on the Draft Online Safety Bill: Joint Committee on the Draft Online Safety Bill (2021). *Written Evidence Submitted by Dr Amy Orben, College Research Fellow at the University of Cambridge*. OSB0131. Available at: https://committees.parliament.uk/writtenevidence/39303/html/ [accessed 3 March 2022].

[50] For further detail, see Joint Committee on the Draft Online Safety Bill (2021). *Written Evidence Submitted by Reset*. OSB01238. Available at: https://committees.parliament.uk/writtenevidence/39303/html/ [accessed 3 March 2022].

[51] Digital, Culture, Media & Sport (2021). *Draft Online Safety Bill*. CP405. London: HMSO, clause 70. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf [accessed 3 March 2022]; also see Chapter 4, clause 85(1).

[52] Digital, Culture, Media & Sport (2021). *Draft Online Safety Bill*. CP405. London: HMSO, clause 74. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf [accessed 3 March 2022]; also Chapter 4, Clause 88.

[53] On the concept of the transparency report, see in particular Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para 27 to 28 and 2.15 to 2.1. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022]; also see Department for Digital, Culture, Media & Sport (2022). *Online Safety Bill*. London. HMSO. (58/2), Chapter 4, Clause 65. Available at: Online Safety Bill publications - Parliamentary Bills - UK Parliament [accessed 29 March 2022].

[54] Gov.UK (2021). *Press Release: Landmark laws to keep children safe, stop racial hate and protect democracy online published*. Available at: https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published [accessed 3 March 2022].

[55] Gov.UK (2021). *Press Release: Landmark laws to keep children safe, stop racial hate and protect democracy online published*. Available at: https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published [accessed 3 March 2022].

[56] Joint Committee on the draft Online Safety Bill (2021). *Report of session 2021-22, 14 December 2021*. HL Paper 129/HC 609, para. 46. Available at: https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/129.pdf [accessed 3 March 2022].

[57] Joint Committee on the Draft Online Safety Bill (2021). *Corrected Oral Evidence, Thursday 9 September 2021*, 9:30am. Available at: https://committees.parliament.uk/oralevidence/2695/html/ [accessed 3 March 2022].

[58] Ada Lovelace Institute and Reset (2020). *Inspecting Algorithms in Social Media Platforms*. Available at: https://www.adalovelaceinstitute.org/report/inspecting-algorithms-in-social-media-platforms/ [accessed 3 March 2022]. Also see Ada Lovelace Institute (2020). *Examining the Black Box: Tools for assessing algorithmic systems*. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/ [accessed 3 March 2022]; Kayser-Bril, N. (2021). AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook. *AlgorithmWatch.* Available at: https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/ [accessed 3 March 2022].

[59] Gov.UK (2021). *Press Release: Landmark laws to keep children safe, stop racial hate*

and protect democracy online published. Available at: https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published [accessed 3 March 2022].

[60] Department for Digital, Culture, Media and Sport (2020). *Online Harms White Paper: Full government response to the consultation*, CM354. London. HMSO, para. 4.18, Box 17. Available from Online Harms White Paper: Full government response to the consultation - GOV.UK (www.gov.uk) [accessed 3 March 2022].

[61] Joint Committee on the Draft Online Safety Bill (2021). *Corrected Oral Evidence, Thursday 14 October*, 9:55am. Available at: https://committees.parliament.uk/oralevidence/2816/html/ [accessed 3 March 2022].

[62] Joint Committee on the Draft Online Safety Bill (2021). *Written Evidence Submitted by Dr Amy Orben, College Research Fellow at the University of Cambridge*. OSB0131. London: HMSO. Available at: https://committees.parliament.uk/writtenevidence/39303/html/ [accessed 3 March 2022].

[63] On this point see Starr, P. (1987). The sociology of official statistics. W. Alonso, P. Starr (eds.). *The Politics of Numbers*. New York: Russell Sage Foundation.

[64] Meredith, J. (2017). Analysing technological affordances of online interaction using conversation analysis. *Journal of Pragmatics*, 115, 42-55; Madell, D., Muncer, S. (2007). Control over social interactions: an important reason for young people's use of the internet and mobile phones for communication? *Cyberpsychology & Behaviour*. 10(1), 137-140; Garcia, A., Jacobs, J. (1999). The eyes of the beholder: understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on Language and Social Interaction*, 32(4), 337-367.

[65] Ditchfield, H. (2019). Behind the screen of Facebook: Identity construction in the rehearsal stage of online interaction. *New Media & Society,* 22(6): 927-943.

[66] Ditchfield, H., Lunt, P. (2021). Reconfiguring synchronicity and sequentiality in online interaction, in A. Kaun (ed.) *Making Time for Digital Lives: Beyond Chronotopia*. London: Rowman & Littlefield; Stephens, K., Panjota, G. (2016). Mobile devices in the classroom: learning motivations predict specific types of multicommunicating behaviours. *Communication Education*, 65(4), 463-479.

[67] Austin, J.L. (1962). *How to do Things with Words*. Oxford: Clarendon Press.

[68] Rose, N. (1988). Calculable minds and manageable individuals. *History of the Human Sciences*, 1(2): 179-200.

[69] See Cline-Cohen, P. (1982). *A Calculating People: The Spread of Numeracy in Early America*. Chicago: University of Chicago Press; and Hopwood, A.G. and Miller, P. (eds.) (1986). *Accounting as Social and Institutional Practice*. Cambridge: Cambridge University Press.

[70] On the concepts of 'mutable' and 'immutable' mobiles, see Latour, B. (1987). *Science in Action*. Milton Keynes: Open University Press.

[71] On numbers and standardization, see Porter, T. (1992). Quantification and the accounting ideal in science. *Social Studies of Science*, 22: 633-652.

[72] See Rose, N. (2000). *Powers of Freedom: Reframing Political Thought*. Cambridge: Cambridge University Press.

[73] See Miller, P. (1992). Accounting expertise and the politics of the product: economic citizenship and modes of corporate governance,' *Accounting, Organisations and Society*, 17: 187-206.


AUDI ALTERAM PARTEM
FSU
FREE SPEECH UNION