CENTRE FOR LAW AND DEMOCRACY

January
2023

Background Paper

Legal Regulation of Platforms to Promote Information as a Public Good

info@law-democracy.org
+1 902 431-3688
www.law-democracy.org
fb.com/CentreForLawAndDemocracy
@Law_Democracy

Table of Contents

Introduction[1]

This Background Paper was prepared by CLD as an input for UNESCO to assist it in preparing for UNESCO's Global Conference on "Internet for Democracy: Regulating Digital Platforms for Information as a Public Good", which took place in Paris from 21-23 February 2023. This followed an announcement by UNESCO Director-General Audrey Azoulay at the 2022 World Press Freedom Conference in Uruguay to the effect that UNESCO would lead on the development of a model global regulatory framework for platforms.

The wider context for this is the massive growth in the dissemination of both illegal and legal but harmful ("lawful but awful") speech online which many commentators have argued is posing a serious threat to the wider information environment. The approach so far has in many cases been largely to leave it up to the companies themselves to address the problems, sometimes while engaging with States and other actors. However, this has so far essentially failed to curb the most serious problems. As a result, States are increasingly taking direct steps to regulate especially those intermediaries which operate as social media platforms. As UNESCO Director-General Audrey Azoulay said in Uruguay, "we cannot leave it to private companies to resolve this existential issue themselves, as their business models will continue to favour engagement and clicks, sometimes at any cost, prioritising sensational content over verified information".

The specific focus of this Paper is on legal developments (as compared, for example, with voluntary initiatives being taken by platforms and other intermediaries or civil society-led initiatives). The first section focuses on international human rights standards relating to freedom of expression. Given the audience, this does not delve into the details of basic freedom of expression standards and instead focuses on issues of particular relevance to regulating the online space and platforms in particular. This section includes parts focusing on restrictions (including the standards which apply to different types of restrictions, such as criminal prohibitions and media regulation), positive obligations, including the responsibility of States to take steps to protect freedom of expression against interference by third parties, the rights to "seek" and "receive" information as part of freedom of expression, the complex balancing issues that arise when these rights conflict with the right to "impart" information and ideas, and some of the implications of these various aspects of freedom of expression for regulating the online space. A final part of this section looks briefly at some repressive laws which clearly represent regulatory overreach, just to signal the illegitimacy of this.

Most of the rest of the Paper looks at legal (i.e. formally obligatory) measures currently being implemented or developed by more democratic jurisdictions. The aim is to highlight ostensibly legitimate or at least legitimately intentioned attempts globally to address the harms caused by online speech (even if, ultimately, these attempts are not deemed to be

[1] This report was authored by Toby Mendel, Executive Director, Centre for Law and Democracy. The Centre for Law and Democracy would like to thank UNESCO for their generous support which made the production and publication of this Background Paper possible.

legitimate). Within these attempts, a special focus is placed on the Digital Services Act (DSA) of the European Union. This is because of the significance of this legislation, given that it covers an important part of the democratic world, that it is extremely detailed and far-reaching in its measures, and that it is actually both finalised and in force (as of 16 November 2022).

The Paper looks at a number of different areas of national (and European Union) regulation. The largest section focuses on the direct legal regulation of illegal content, the primary aim of the DSA and most of the other initiatives. That section is, in turn, broken down into parts looking at the core way the system works (such as by protecting intermediaries against liability unless they fail to take certain actions), the actors and content covered (respectively), how the obligation to act is triggered (or how intermediaries are supposed to become aware of illegal content), the measures required to be taken, complaints systems for users whose content has been affected, institutional measures and other issues.

This is followed by sections looking at the legal regulation of legal (i.e. as opposed to illegal) content, engaging regulatory, co-regulatory and self-regulatory bodies, and then indirect measures to support content regulation, with a strong focus on transparency although also looking at issues such as human rights due diligence and safety obligations. A final section provides a human rights analysis of the various systems canvassed, along with recommendations.

1. Wider Freedom of Expression Considerations

This part of the Paper outlines various freedom of expression considerations which are of particular importance to the regulation of platforms to promote information as a public good.

1.1. Restrictions

The right to freedom of expression, protected in Article 19 of the *Universal Declaration of Human Rights* (UDHR),[2] and the same article in the *International Covenant on Civil and Political Rights* (ICCPR),[3] is a complex right. On the one hand, although freedom of expression is not an absolute right, it strictly limits the power of the State to impose restrictions on it, so-called negative obligations (because they set out what States cannot do). Thus, Article 19(3) of the ICCPR states:

> The exercise of the right to freedom of expression] carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:
>
> (a) For respect of the rights or reputations of others;

[2] UN General Assembly Resolution 217A (III), 10 December 1948.
[3] UN General Assembly Resolution 2200A (XXI), 16 December 1966, in force 23 March 1976, http://www2.ohchr.org/english/law/ccpr.htm.

> (b) For the protection of national security or of public order (ordre public), or of public health or morals.[4]

This effectively translates into a three-part test for restrictions on freedom of expression, namely that any restriction must:

1. be provided by law;
2. protect one of the (legitimate) interests listed there; and
3. be necessary for the protection of that interest.

This test is applied rigorously in all cases involving restrictions on freedom of expression which are decided by international and regional human rights courts and oversight bodies.[5]

Any obligations States impose on platforms to restrict (limit) content must meet this three-part test, including that they can be justified as necessary to protect a legitimate interest. However, international law recognises that there are important differences between different mediums of communication and that standards which govern one medium may not apply in the same way to others. For example, it is one thing to impose criminal liability on an individual for speech which reflects intolerance (hate speech), and quite another to impose professional standards on licensed broadcasters as an administrative law matter in relation to the same type of speech. International law sets out clear and narrow parameters for the former, as reflected in Article 20(2) of the ICCPR,[6] while democracies around the world, as well as self-regulatory systems run by broadcasters themselves, have adopted much higher standards regarding what sorts of intolerant speech broadcasters may not disseminate, often imposed via mandatory professional codes of conduct. The same analysis applies to content disseminated over platforms, as a very unique medium of communication.

The implications of this in the case of inaccurate statements are important for our analysis. It is now clear that it is not legitimate to impose general bans on the dissemination of false or inaccurate statements. Thus, in their 2017 Joint Declaration, the special international mandates on freedom of expression at the United Nations (UN), Organization for Security and Co-operation in Europe (OSCE), Organization of American States (OAS) and African Commission on Human and Peoples' Rights stated:

> General prohibitions on the dissemination of information based on vague and ambiguous ideas, including "false news" or "non-objective information", are incompatible with international standards for restrictions on freedom of expression, as set out in paragraph 1(a), and should be abolished.[7]

[4] Very similar standards apply under the three regional systems for the protection of human rights in Africa, the Americas and Europe.
[5] See UN Human Rights Committee, General Comment No. 34, Article 19: Freedoms of opinion and expression, 12 September 2011, para. 22. Available in all six UN languages at: http://undocs.org/ccpr/c/gc/34.
[6] This states: "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."
[7] Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda, 3 March 2017, para. 2(a), https://www.law-democracy.org/live/legal-work/standard-setting/. The mandates have adopted a Joint Declaration every year since 1999.

This rule is respected in many countries. All countries have rules penalising the making of false statements in specific contexts – such as lying in court (perjury) or making false statements which harm the reputation of a third party (defamation) – but democracies otherwise allow even the international telling of lies.

On the other hand, professional media codes of conduct, whether adopted as part of a regulatory, co-regulatory or self-regulatory system, normally impose a much higher standard, requiring media outlets to report with "due accuracy" or some such formulation, and to correct any mistakes that are made.[8] Whereas perjury and making defamatory statements may, respectively, lead to criminal and civil liability, any sanction for failing to meet the standard of "due accuracy" is normally fairly light.[9] Of course all of these different rules must always meet the standards of the three-part test for restrictions on freedom of expression. But the necessity part of the analysis plays out differently in the context of imposing criminal penalties on an individual and requiring the media to respect professional administrative standards.

Platforms are very different both from individual modes of expression and from the media. Neither of the fairly well-established standards outlined above would work (or be legitimate) for platforms. However, requirements for platforms to take certain measures to "address" disinformation, depending on the specific nature of those measures and the nature of the operations of the platform, might pass muster as restrictions on freedom of expression depending on how they were designed. For example, requiring platforms to label or deprioritise disinformation is very different from requiring them to remove or block access to it, let alone suspending a user's account for uploading it. A key issue with any such requirement would undoubtedly be the practicality of applying it at a technical level, as well as the risks such an approach posed to legitimate speech. Until now, we have seen relatively limited standard-setting in this area by international courts and what does exist usually focuses on fairly specific situations. As such, there is potentially significant room for development of our understanding of freedom of expression in this area.

1.2. Positive Obligations

Beyond limiting the power of States to restrict speech, international guarantees of freedom of expression also call on States to take steps or put in place measures to protect the free flow of information and ideas in society, so-called positive obligations (because they require rather than preclude State action). Thus, in the case of *Özgür Gündem v. Turkey*, the applicant newspaper was subjected to such serious attacks and harassment that it was eventually forced to close. The European Court of Human Rights held Turkey directly responsible for

[8] See, as just one example, the Broadcasting Code of the United Kingdom broadcast regulator, Ofcom, Section five: Due impartiality and due accuracy, https://www.ofcom.org.uk/tv-radio-and-on-demand/broadcast-codes/broadcast-code/section-five-due-impartiality-accuracy.
[9] Although repeated and blatant breaches, especially made with an ulterior motive, may lead to heavier sanctions.

certain acts of harassment. It also recognised that, under certain circumstances, States have a positive obligation to protect freedom of expression, stating:

> The Court recalls the key importance of freedom of expression as one of the preconditions for a functioning democracy. Genuine, effective exercise of this freedom does not depend merely on the State's duty not to interfere, but may require positive measures of protection, even in the sphere of relations between individuals. In determining whether or not a positive obligation exists, regard must be had to the fair balance that has to be struck between the general interest of the community and the interests of the individual, the search for which is inherent throughout the Convention. The scope of this obligation will inevitably vary, having regard to the diversity of situations obtaining in Contracting States, the difficulties involved in policing modern societies and the choices which must be made in terms of priorities and resources. Nor must such an obligation be interpreted in such a way as to impose an impossible or disproportionate burden on the authorities. [references omitted][10]

As this quotation makes clear, in some cases States have a positive obligation to intervene to prevent private third parties from unduly interfering with the freedom of expression rights of individuals. In that case, the issue was safety – i.e. the obligation of the State to provide protection to a media outlet which was being attacked by private third parties (as well as State actors) – but such obligations have also been found in areas such as the right to information,[11] protection of confidential journalistic sources,[12] the promotion of media diversity,[13] and even providing legal aid to defendants in private defamation cases in certain circumstances.[14]

Online intermediaries, and especially the main social media platforms, play an increasingly central role in facilitating, and managing, modern communications of all sorts. As such, it seems obvious that, should they fail to ensure robust protection for freedom of expression, including the right to access information, this would be an obvious candidate for positive obligations on States to intervene to protect the free flow of information and ideas in society. The quotation above also makes it clear that the scope of positive obligations under international law depend on an assessment of the balance between individual and collective or community interests. This balancing is often front and centre in the digital communications space, as elaborated on more below.

1.3. The Rights to Seek and Receive

[10] *Özgür Gündem v. Turkey*, 16 March 2000, Application No. 23144/93, para. 43.
[11] See, for example, *Claude Reyes and Others v. Chile*, 19 September 2006, Series C., No. 151, (Inter-American Court of Human Rights), http://www.corteidh.or.cr/docs/casos/articulos/seriec_151_ing.pdf.
[12] See, for example, *Goodwin v. the United Kingdom*, 27 March 1996, Application No. 17488/90 (European Court of Human Rights).
[13] See, for example, the 2007 Joint Declaration on Promoting Diversity in the Broadcast Media of the special international mandates on freedom of expression, 12 December 2007, https://www.osce.org/fom/66176?page=1. See also *Centro Europa 7 S.R.L. and Di Stefano v. Italy*, 7 June 2012, Application No. 38433/09, paras. 129-134 (European Court of Human Rights).
[14] See *Steel and Morris v. the United Kingdom*, 15 February 2005, Application No. 68416/01, para. 95 (European Court of Human Rights).

For most people, the primary association of the right to freedom of expression is with the right to express oneself or to "impart" information and ideas. Clearly this is absolutely central to the right. However, international guarantees of freedom of expression are not limited to the speaker. Thus, Article 19 of both the UDHR and the ICCPR protect the rights to "seek, receive and impart" information and ideas. While the last of these refers to the right to speak, the other two focus on the recipient of the information, or what we might call the rights of the listener.

The Inter-American Court of Human Rights has so far elaborated more clearly on the "seek" and "receive" aspects of freedom of expression than any other international human rights court.[15] In an Advisory Opinion where the Court was called upon to assess the legitimacy of a mandatory licensing system for journalists, the Court elaborated on dual nature of freedom of expression:

> [W]hen an individual's freedom of expression is unlawfully restricted, it is not only the right of that individual that is being violated, but also the right of all others to "receive" information and ideas. The right protected by Article 13 consequently has a special scope and character, which are evidenced by the dual aspect of freedom of expression. It requires, on the one hand, that no one be arbitrarily limited or impeded in expressing his own thoughts. In that sense, it is a right that belongs to each individual. Its second aspect, on the other hand, implies a collective right to receive any information whatsoever and to have access to the thoughts expressed by others…. In its social dimension, freedom of expression is a means for the interchange of ideas and information among human beings and for mass communication. It includes the right of each person to seek to communicate his own views to others, as well as the right to receive opinions and news from others. For the average citizen it is just as important to know the opinions of others or to have access to information generally as is the very right to impart his own opinions.[16] [references omitted]

Taken together, the rights to both impart and to seek and receive information create a wide vision for freedom of expression as protecting the free flow of information and ideas in society. Many of the positive obligations of States to protect freedom of expression noted above are ultimately based on the idea of promoting the rights to seek and receive information and ideas. Thus, this is true of the right to information,[17] the protection of confidential journalistic sources[18] and certainly the promotion of media diversity. The latter, especially when considered in light of the prohibition on discrimination in the enjoyment of

---

[15] Article 10 of the *European Convention on Human Rights*, adopted 4 November 1950, in force 3 September 1953, protecting freedom of expression, refers only to the rights to "receive and impart" information and ideas and not to "seek" them. It is not clear whether and, if so how far, this makes a difference in terms of this issue.

[16] *Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism*, Advisory Opinion OC-5/85 of 13 November 1985, Series A, No. 5, paras. 30-2.

[17] In the leading case on this issue before the Inter-American Court of Human Rights, *Claude Reyes and Others v. Chile*, 19 September 2006, Series C, No. 151, the Court very clearly and explicitly linked this aspect of freedom of expression to the rights to seek and receive information and ideas. See para. 77.

[18] Contrary to what is sometimes said about this right, although it does apply primarily to journalists and the media, it is not limited to those actors and the purpose is not to grant them, *per se*, special rights but to protect their ability to act as intermediaries between sources and the public, which has a right to hear what the sources want to tell them. As the European Court stated in a leading case on this issue, absent source protection, "the ability of the press to provide accurate and reliable information [to the public] may be adversely affected". *Goodwin v. the United Kingdom*, note 12, para. 39.

rights (for example as protected in Article 2 of the ICCPR), creates obligations on States to ensure access for all groups in society to the media. Thus, the OAS Special Rapporteur for Freedom of Expression has noted:

> It is therefore clear that the regulation of radio broadcasters must aim to overcome the preexisting inequalities in access to the media, which include, for example, that of economically disadvantaged sectors of society. In this sense, States must not only refrain from discriminating against these sectors, but also promote proactive public policies for social inclusion.[19]

These comments were made in the context of broad acceptance of the necessity for States to regulate broadcasting. But the core thrust of this idea is that everyone, whatever their position in society, has the right to access a wide diversity of information and ideas. In essence, individuals have a right to hear not just the views of those who happen to own media outlets but to a media system which reflects all views and perspective held in society without discrimination.

Often, the rights to seek and receive, and then to impart, information and ideas are aligned, as in the cases of protection of sources and the right to information. However, they can also come into conflict. Thus, measures to prevent undue concentration of media ownership, which are widely accepted not only as legitimate but indeed necessary,[20] represent both restrictions on the rights of speakers (specifically owners seeking to expand their power to speak) and protection of the rights of listeners (i.e. to receive a diversity of information and ideas).

Such cases raise very complex jurisprudential issues, such as what standard or test to apply to assess which set of rights should dominate. This cannot be done through a simple application of the three-part test for restrictions, since the essence of this is to create a strong presumption in favour of freedom of expression as against other social interests whereas here we have a conflict between two opposing freedom of expression interests. Put differently, if courts applied the three-part test in such cases, the result would be very different, even based on identical facts, depending on who brought the case (the speaker or the listener), which is clearly not a credible outcome. Courts have fashioned strong analytical approaches for dealing with other cases where rights come into conflict, such as where privacy conflicts with freedom of expression.[21] In such cases, they have essentially introduced special considerations under the third part of the test for restrictions (since the requirement for restrictions to be set out in law clearly continues to apply and the fact that the law aims to

[19] *Freedom of Expression Standards for Free and Inclusive Broadcasting*, 30 December 2009, para. 35, https://www.oas.org/en/iachr/expression/docs/publications/Broadcasting%20and%20freedom%20of%20expresion%20FINAL%20PORTADA.pdf. The same report devotes quite a lot of attention to States' obligations in terms of community broadcasting, which extends to reserving spectrum for its use and putting in place fair and simple licensing regimes for this broadcasting sector. See paras. 96-113.

[20] See, for example, Toby Mendel, Ángel Garcia Castillejo and Gustavo Gómez, *Concentration of Ownership and Freedom of Expression: Global Standards and Implications for the Americas* (2017, Paris, UNESCO), https://unesdoc.unesco.org/ark:/48223/pf0000248091.

[21] See, for example, the two cases, *Von Hannover v. Germany*, 24 June 2004, Application No. 59320/00 and *Von Hannover v. Germany (No. 2)*, 7 February 2012, Applications Nos. 40660/08 and 60641/08, before the European Court of Human Rights.

protect a human right automatically satisfies the second part of the test). Unfortunately, international courts have not yet provided a clear jurisprudential analysis for resolving conflicts between the freedom of expression rights of the speaker and listener.

1.4. Implications Online

It is increasingly recognised that States have a positive obligation to promote universal access to the Internet. Thus, as far back as 2011, in their Joint Declaration on Freedom of Expression and the Internet, the special international mandates on freedom of expression indicated: "Giving effect to the right to freedom of expression imposes an obligation on States to promote universal access to the Internet."[22] Beyond this, there is a growing recognition that "zero-rating" schemes, whereby Internet access providers effectively give users free access to certain services, content or applications, usually prominently featuring Facebook, while they have to pay to access other services (or, if their plan does not include data, they simply cannot access those other services), also represent a breach of the right to freedom of expression, partly because they represent a breach of net neutrality and partly because they effectively deny, on a discriminatory basis, access of some users to the whole Internet.[23]

There is something of a debate globally about standards for liability of intermediaries for third party content, although it is recognised that different standards are appropriate for different types of intermediaries. It is widely agreed that access providers, platforms and similar intermediaries should not generally be required to monitor the content flowing through or hosted on their systems. This is reflected in Principle 6 of the Council of Europe's Declaration on Freedom of Communication on the Internet:

> Member States should not impose on service providers a general obligation to monitor content on the Internet to which they give access, that they transmit or store, nor that of actively seeking facts or circumstances indicating illegal activity.[24]

There is, however, less agreement as to when intermediaries should become liable for third party content. In their 2011 Joint Declaration, the special mandates called for broad protection against liability for those providing "technical Internet services", unless they refused to obey a court order to remove the content, and then added:

> Consideration should be given to insulating fully other intermediaries, including those mentioned in the preamble, from liability for content generated by others under the same conditions as in paragraph 2(a). At a minimum, intermediaries should not be required to monitor user-generated content and should not be subject to extrajudicial content takedown rules which fail to provide sufficient protection for freedom of expression (which is the case with many of the 'notice and takedown' rules currently being applied).[25]

[22] Adopted 1 June 2011, para. 6(a), http://www.law-democracy.org/wp-content/uploads/2010/07/11.06.Joint-Declaration.Internet.pdf.
[23] See, for example, Centre for Law and Democracy, *Colombia: Amicus Brief in Net Neutrality Case*, February 2022, https://www.law-democracy.org/live/legal-work/litigation/.
[24] Adopted by the Committee of Ministers on 28 May 2003, https://rm.coe.int/16805dfbd5.
[25] Note 22, para. 2(b).

And strong versions of this exist in some countries, such as is reflected in section 230 of the United States' Communications Act.[26]

On the other hand, in Europe, the notice and takedown rules that the quotation above questions are widely applied to services other than those providing "mere conduit" for content[27] and these rules have largely been endorsed by the European Court of Human Rights.[28]

Beyond liability, the Inter-American Court of Human Rights has, in a criminal defamation case, suggested that State obligations to promote pluralism go beyond just the media and apply to the information space more generally, stating:

> Given the importance of freedom of thought and expression in a democratic society … the State must not only minimize restrictions on the dissemination of information, but also extend equity rules, to the greatest possible extent, to the participation in the public debate of different types of information, fostering informative pluralism.[29]

Expanding the idea of pluralism to the Internet has potentially very significant implications. In many ways pluralism is strongly wired into the very nature of the Internet, which has created unprecedented opportunities for ordinary people to speak to the world, as it were, essentially at very low cost. At the same time, it is now clear that a variety of factors, including the business models of some of the key social media platforms, along with human nature, has created undue dominance of certain forms of speech, in particular disinformation, over others. As the UN Special Rapporteur on Freedom of Expression wrote in 2021:

> False information is amplified by algorithms and business models that are designed to promote sensational content that keep users engaged on platforms. Disinformation thrives in an online environment that encourages amplification while reducing accessibility to plural and diverse sources of information.[30]

In a report just two years earlier on a similar subject titled *Guide to guarantee freedom of expression regarding deliberate disinformation in electoral contexts*, the OAS Special Rapporteur for Freedom of Expression highlighted a number of factors driving the prevalence of disinformation online, including political polarisation, social media bubbles, the nature of online advertising and the apparently natural emotional attraction of this sort of content.[31]

---

[26] 47 U.S. Code § 230, 1934 Communications Act.
[27] See, for example, the so-called E-Commerce Directive 2000/31/EC of the European Union, 8 June 2000, Articles 12-14, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32000L0031.
[28] For example, in the case of *Delfi AS v. Estonia*, Application No. 64569/09, 16 June 2015, which involved the liability of an online news platform for user-generated content, a Grand Chamber of the European Court noted with approval the fact that the platform had removed the content as soon as it was notified of its illegality. See para. 76. See also *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v. Hungary*, Application No. 22947/13, 2 February 2016.
[29] *Kimel v. Argentina*, 2 May 2008, Series C, No. 177, para. 57.
[30] *Disinformation and freedom of opinion and expression*, 13 April 2021, para. 16, https://www.ohchr.org/en/documents/thematic-reports/ahrc4725-disinformation-and-freedom-opinion-and-expression-report.
[31] October 2019, pp. 14-17, https://www.oas.org/en/iachr/expression/publications/Guia_Desinformacion_VF%20ENG.pdf.

Both Special Rapporteurs also highlighted the corrosive impact of disinformation. The UN Rapporteur wrote:

> Interacting with political, social and economic grievances in the real world, disinformation online can have serious consequences for democracy and human rights, as recent elections, the response to the coronavirus disease (COVID-19) pandemic and attacks on minority groups have shown. It is politically polarizing, hinders people from meaningfully exercising their human rights and destroys their trust in Governments and institutions.[32]

For his part, the OAS Rapporteur, consistently with the focus of that report on electoral contexts, noted:

> [I]t seems clear that the deliberate spread of false information impoverishes the public debate and makes it harder for citizens to exercise their right to receive information from various sources, and in the end it is an obstacle to participating in Democratic decisions.[33]

As noted above, in many cases promoting diversity involves a trade-off between different freedom of expression values. Certainly any regulatory measures to counter disinformation online would fall into this category and, in balancing the values, it needs to be taken into account that much disinformation is, ultimately, political speech which receives a very high degree of protection under the right to freedom of expression. At the same time, as the Special Rapporteurs noted, disinformation can very seriously undermine human rights and even participation in democratic decision-making. Ultimately, disinformation can erode our core relationship with information, such that we no longer have confidence that we know what is true and what is not, which is somehow at the very core of what freedom of expression stands for.

Different forms of online harassment, sometimes referred to as mal-information, can have a direct impact on the freedom of expression of speakers to the extent that this behaviour intimidates its targets from exercising this right. This is especially the case where mal-information is specifically directed at purveyors of public interest information, such as journalists, with the goal of shutting them up. Furthermore, in many cases these instances of mal-information are discriminatory in nature, being directed specifically at identifiable groups. Thus, the phenomenon of online violence against journalists is heavily biased towards women, with a recent UNESCO study showing that 73% of female journalists had experienced online violence. One of the conclusions of this report was that there was a prevalence of: "Constant moderate-low volume abuse and harassment that burns slowly but can be cumulatively devastating."[34]

Whereas lower volumes of certain types of speech do not create a sufficient level of harm to pass the necessity part of the test for restrictions, as the volume of such speech increases, the level of harm increases as well, and the necessity calculation starts to change. On this theory, while it would still not be legitimate to sanction the original author of each discreet statement,

---

[32] Note 30, para. 2.
[33] Note 31, p. 17.
[34] Julie Posetti, Nabeelah Shabbir, Diana Maynard, Kalina Bontcheva and Nermine Aboulez, *The Chilling: Global trends in online violence against women journalists*, April 2021, pp. 11-12, https://unesdoc.unesco.org/ark:/48223/pf0000377223.

on the basis that he or she has not alone caused harm, measures, even of a mandatory nature, to address the harm which the cumulative dissemination of such statements can cause, and which is the result of the systemic dissemination of these statements by corporate actors, may be legitimate, of course as always taking into account the specific nature of those measures and the impact they have on freedom of expression. Cyber bullying laws, which seek to respond to the undoubtedly cumulatively harmful nature of what can turn into continuous, large-scale online bullying, have sometimes failed to successfully negotiate this boundary between individual (relatively harmless) statements and the cumulative impact of collective such statements (i.e. by banning the former).[35] Of course it is not always easy to find a systemic (intermediary-level) solution to such problems.

Where these situations represent clashes between different freedom of expression interests, that also needs to be taken into account. Harassment of journalists not only pits their right to freedom of expression against that of those who are harassing them, but also the right of others to receive the public interest content those journalists disseminate. In the case of online violence against female journalists and other public interest speakers, to the extent that this results in them withdrawing from disseminating public interest content, that is not only (also) an attack on the right of those who consume their content to seek and receive information, but it also has the effect of perpetuating social discrimination in the information space. Of course extreme care must also be taken here not to skew social debate in another way, namely by preventing robust criticism of various points of view.

Looked at from the specific perspective of dis- and misinformation, where citizens are so overwhelmed by these forms of speech that they find it difficult to discern what is actually true, the harm to both freedom of expression and the social values it underpins including, ultimately, democracy, are obvious. Indeed, a powerful argument in favour of freedom of expression is precisely that it allows the truth to prevail, for example over the interests of powerful social actors. Addressing threats to the ability of free speech to privilege, ultimately, the truth can thus be seen as part of the defence of free speech.

1.5. Regulatory Overreach

The previous section both highlighted limits on restrictions on online speech and pointed to some justifications for potentially new forms of structural regulation of platforms based on the changes brought about by digital speech and the new forms of communications it has engendered. Unfortunately, it has become common for countries to take advantage of this fluid situation to pass repressive laws unduly limiting online freedom of expression, often as part of a wider attempt to control social and particularly political debate. While these can address any number of content issues, there is a strong focus on banning inaccurate statements (disinformation). As the UN Special Rapporteur on freedom of expression noted

[35] Even more structural measures can fall foul of the right to freedom of expression. See, for example, a case from Nova Scotia in which a court struck down an administrative scheme to limit (as opposed to a criminal proscription on) cyberbullying: https://www.cbc.ca/news/canada/nova-scotia/cyberbullying-law-struck-down-1.3360612.

about responses to disinformation in her 2021 report on *Disinformation and freedom of opinion and expression*:

> State responses have often been problematic and heavy handed and had a detrimental impact on human rights.[36]

This is not an entirely new phenomenon. As the Centre for Law and Democracy wrote in its *Submission on an Annual Thematic Report on Disinformation* to the UN Special Rapporteur for Freedom of Expression:

> Laws prohibiting disinformation are not a new invention. France, for example, still has a prohibition on spreading false news in its Freedom of the Press Law, which dates from 1881.[37]

But there has been a rash of adoption of such laws recently, which has increased since the advent of the COVID-19 pandemic.

To give just a few recent examples of broad prohibitions on disinformation, Singapore's Protection from Online Falsehoods and Manipulation Act 2019 prohibits making false statements of facts which result in a number of very broadly defined harms, such as: prejudicing the security of Singapore; prejudicing public health, safety and tranquillity; influencing the outcome of an election; inciting feelings of enmity or ill-will between groups of persons; or diminishing public confidence in the government. The penalty for doing so is up to five years' imprisonment and/or a fine.[38] While this is not quite an absolute ban on any inaccurate statement, the consequences are defined very broadly indeed and refer to issues which go far beyond the interests which Article 19(3) of the ICCPR accepts as grounds to restrict freedom of expression (including diminishing confidence in the government or prejudicing public tranquillity). Ironically, the German NetzDG Law has been cited as one of the justifications for the Singaporean law, even though the former does not specifically reference false statements (even if some of the many crimes it does cover do include elements of falsity, among other things).[39]

Malaysia's Anti-Fake News Act 2018 was even broader, banning any malicious act to create or disseminate fake news, defined as any information which is wholly or partially false. However, the Act was only in force for a few months before being repealed in August 2018.[40]

---

[36] Note 30, para. 3.

[37] P. 4, https://www.law-democracy.org/live/un-special-rapporteur-on-freedom-of-expression-submission-on-disinformation/. See generally pp. 4-6.

[38] Protection from Online Falsehoods and Manipulation Act 2019, No. 18 of 2019, section 7, https://sso.agc.gov.sg/Acts-Supp/18-2019.

[39] See, for example, Jacob Mchangama and Joelle Fiss, *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship*, Justicia, November 2019, p. 9, https://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf.

[40] Act No. 803, ss. 2 and 4(1), https://www.cljlaw.com/files/bills/pdf/2018/MY_FS_BIL_2018_06.pdf.

In August 2021, Cuba enacted new regulations criminalising the dissemination of "false" and "offensive" information online,[41] another clearly illegitimate rule.

This is just a very small selection from among a broad range of restrictions adopted by countries around the world which prohibit the dissemination (and often merely the creation or holding of) digital content far more broadly than Article 19(3) of the ICCPR allows.

2. Legal Regulation of Illegal Content

This section focuses on legal measures which have been put in place in democracies, including both more established, or Western, democracies and newer democracies, to address through structural or systemic measures speech which is criminally prohibited, although in some cases the rules extend beyond that to civil liability as well. We have reviewed rules adopted in Australia, Canada, the European Union, Germany, Spain and the United Kingdom for the first group, and Brazil, India and South Africa, for the second group.

The section covers both laws (regulations and so on) which have been passed and, in a few cases, legislative proposals/draft laws. This space is evolving very rapidly around the world and it is important to cover emerging rules as well as settled ones. Even the European Union's DSA[42] only entered into force very recently, on 16 November 2022, although it was originally proposed in December 2020.[43]

The section starts by describing the core manner in which these systems work, whether this essentially focuses on the scope of exemptions from liability or imposing additional obligations on different actors. It then canvases which actors are covered by the system, followed by the types of content covered. Next, the key issue of what triggers the responsibilities of intermediaries to act – such as upon being provided with notice – are reviewed. This is followed by a review of the different measures that are required to be taken against the content, and then of the sorts of complaints systems which must be made available for users. Some key institutional features of these systems are then reviewed, albeit only inasmuch as they are relevant to a proper consideration of the way rights play out under these systems. Finally, this section reviews other issues which are relevant.

2.1. How the System Works

Most of these systems have at their core the idea of imposing liability on certain intermediaries in case they fail to take measures to address certain types of content or,

[41] Decreto-Ley No. 35 De las Telecomunicaciones, las Tecnologías de la Información y la Comunicación y el Uso del Espectro Radioeléctrico, ***GOC-2021-759-O95***, 17 August 2021, Article 15(e), available in Spanish at: http://media.cubadebate.cu/wp-content/uploads/2021/08/goc-2021-o92-comprimido.pdf.
[42] 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, COM(2020) 825 final, 15 December 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0825&from=en.
[43] See European Commission, The Digital Services Act package, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

alternatively, via imposing responsibility on them to take certain steps to address certain types of content (and liability on them if they fail to do so).

The DSA largely tracks the much earlier EU (2000) E-Commerce Directive[44] in the sense of exempting certain intermediaries from liability under certain conditions but with a few important differences. One is that the DSA is a regulation, as opposed to a directive, which means that it has direct binding force of law in Member States, while States are given some discretion as to how to implement or transpose a directive, albeit that must be done effectively in the sense of achieving the objectives set by the directive. Another is that the DSA contains far more detailed rules and structures around liability and other measures.

The very first provision of the DSA, Article 1(1)(a), indicates that it provides "conditional exemption from liability of providers of intermediary services". It maintains the E-Commerce Directive's categorisation of intermediary services into "mere conduit", "caching" and "hosting". The former are protected against liability unless they intervene in the content. Caching services are also largely protected against liability unless they engage with the content, although they must disable access upon obtaining "actual knowledge" that the information at the initial source has been removed or disabled or that a court has ordered this. Hosting services, on the other hand, lose their protection against liability if they do not act expeditiously to "remove or to disable access" once they obtain "actual knowledge" effectively that the content is illegal (Articles 3-5). Article 42 does also provide for penalties for failure to respect the provisions of the DSA, which penalties shall be "effective, proportionate and dissuasive", although it is unclear if this applies the primary obligation to remove or disable access, the many other obligations to put in place various measures or take various actions or both.

The German Network Enforcement Act (NetzDG), adopted in 2017,[45] is essentially cast as a system to handle complaints, with the key operative provision, Article 1(3), being titled "Handling of complaints about unlawful content". This requires all providers of social networks to maintain an "effective and transparent procedure for handling complaints about unlawful content". However, this procedure includes a requirement to remove or block unlawful content. Article 1(4)(1)(2) provides for penalties for anyone who breaches Article 1(3)(1), with fines of up to EUR five million, although a judicial decision to the effect that the content is actually unlawful is required for this purpose (Article 1(4)(5)).

[44] Note 27.

[45] Act to Improve Enforcement of the Law in Social Networks, 12 July 2017, https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2. The analysis here is largely based on the original version of NetzDG although it has been amended twice, first by the Act to Combat Right-Wing Extremism and Hate Crime, 3 April 2021, and second by the Act Amending the Network Enforcement Act, 28 June 2021. The second set of amendments have been described by the Library of Congress as follows: "The amendment aims to increase the information content and comparability of social media providers' transparency reports and improve the user-friendliness of the reporting channels for complaints about unlawful content." See https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/.

The Indian Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (Indian Rules)[46] were adopted under the Information Technology Act, 2000.[47] They essentially remove the general exemption from liability intermediaries enjoy under section 79(1) of the Act where an intermediary fails to observe the stipulations set out in the Rules (section 7 of the Rules). The Rules also set out a very detailed regime for news publishers and on-demand services (referred to as "publishers of online curated content") in Part III, along with a separate regime of enforcement for them (see below under Engaging Regulatory, Co-regulatory and Self-regulatory Bodies).

In Brazil, the flagship Law No. 12.965, Establishing the principles, guarantees, rights and obligations for the use of Internet in Brazil,[48] popularly known as the Marco Civil, adopted in 2014, set out the basic framework of rules governing the Internet, including such things as liability of intermediaries, the principle of net neutrality, the right of access to the Internet and respect for privacy and data protection principles online. It established that Internet access providers could not be held liable in civil damages for content generated by third parties (Article 18) and that other intermediaries could only be held civilly liable for such content if, following a court order, they did not take steps to block access to that content (Article 19).

In June 2020, the Brazilian Senate passed the Law of Freedom, Responsibility and Transparency on the Internet (Fake News Law, as it was called informally and even on the official Brazilian Senate website).[49] However, the Chamber of Deputies (lower house) refused to fast-track the bill and, as a result, it was not passed before the general elections in October 2022. It is reviewed here as an example of a proposed scheme for regulation of platforms. It specifically indicated that its provisions "must consider" the principles of the Marco Civil (Article 2). At the same time, it provided for various sanctions – including warnings, fines, temporary suspension of activities and ultimately termination of activities (the latter two to be applied only after the first two had already been applied within 12 months of the offence under consideration) – for breach of its provisions (Article 28).

In Canada, the system remains at the stage of a general proposal, in the form of a Discussion Guide[50] and Technical Paper,[51] while draft legislation, originally promised in the fall of 2021, has still not been made public.[52] Interestingly, unlike many other systems, the Technical

[46] Adopted 25 February 2021, https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf.
[47] Act No. 21 of 2000, 9 June 2000, https://prsindia.org/files/bills_acts/bills_parliament/2021/IT%20Act,%202000.pdf.
[48] Adopted 24 April 2014, available in Portuguese and English at: https://publicknowledge.org/policy/marco-civil-english-version/.
[49] Projeto de Lei n° 2630, de 2020, 30 June 2020, https://www25.senado.leg.br/web/atividade/materias/-/materia/141944.
[50] 29 July 2021, https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/discussion-guide.html.
[51] 29 July 2021, https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/technical-paper.html.
[52] Although these two documents, taken together, only provide a rough outline of the proposed system, it is reviewed here due to some rather innovative features as compared to other proposals or systems.

Paper does not work via the approach of lifting immunity. Rather, regulated entities are required to comply with orders issued by various oversight bodies and may be assessed an "administrative monetary penalty" or AMP of up to three percent of their gross global revenue or CAD ten million, whichever is higher, for failing to do so. These may include orders to takedown content or more general compliance orders "to do any act or thing, or refrain from doing anything necessary to ensure compliance" with the rules (clauses 80, 94 and 108). Thus, while regulated entities are required to address regulated content, they are not required to make a correct decision as to whether or not to take it down (i.e. the system does not incentivise entities to be overinclusive in terms of taking content down, as most other systems do).

2.2. Actors Covered

As noted above, the DSA applies to "mere conduit", "caching" and "hosting" services. The former transmit, in a communication network, information provided by a "recipient of the service" (user), or provide access to a communication network. A caching service also consists of transmission but where this involves the "automatic, intermediate and temporary storage" of the information, for purposes of facilitating the operation of the system. Finally, a hosting service stores information provided by and at the request of a user, and thus covers social media platforms.

All intermediaries are covered by the primary DSA obligation to "remove or to disable access". Section 3 (Articles 16-24) establishes various additional obligations for "online platforms" – defined as providers of hosting services which, at the request of users, store and disseminate to the public information – such as in relation to internal complaints, flaggers and reporting. This section does not apply to "micro or small enterprises" as defined in "the Annex to Recommendation 2003/361/EC", which defines the latter as enterprises which employ "fewer than 50 persons and whose annual turnover and/or annual balance sheet total does not exceed EUR 10 million".[53] Section 4 (Articles 25-33) applies to "very large online platforms", defined as those with at least 45 million users in the EU.[54] It imposes additional obligations on these actors, for example in relation to risk assessment and mitigation, independent audits, transparency around recommender systems and the appointment of compliance officers.

NetzDG is designed for social media platforms (social networks) "which are designed to enable users to share any content with other users or to make such content available to the public" and specifically excludes intermediaries offering journalistic content or which are designed to enable individual communications or the "dissemination of specific content". It

[53] Commission Recommendation 2003/361/EC of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises, Annex, Article 2(2), https://service.betterregulation.com/document/175768. Technically micro-enterprises are even smaller but since the DSA does not distinguish between small and micro, this is not operationally relevant.
[54] We were not able to verify this, but this would appear to encompass only Facebook at the moment.

also only applies to intermediaries which have at least two million registered users in Germany (Article 1(1)(1) and (2)).

The Indian Rules broadly apply to intermediaries (see section 3), which term is not defined. However, the temporary storage of information for onward transmission to another computer resource which does not involve "any human, automated or algorithmic editorial control" is not covered (section 3(1)(e)).

There are additional obligations for significant social media intermediaries (see section 4), which are social media intermediaries ("an intermediary which primarily or solely enables online interaction between two or more users and allows them to create, upload, share, disseminate, modify or access information using its services", section 1(w)) and which have a number of users above a threshold notified by the Central Government (section 1(v)). The responsible minister may also require any intermediary to comply with the more onerous obligations of a significant social media intermediary where that intermediary permits the publication of information which may "create a material risk of harm to the sovereignty and integrity of India, security of the State, friendly relations with foreign States or public order" (section 6).

The Indian Rules also put in place a very specific and much more onerous regime for publishers of news and current affairs content (news publishers) and publishers of online curated content (on-demand services) which operate or target India (section 8). The former covers any online paper or similar entity which publishes news and current affairs but does not include newspapers, "replica e-papers of the newspaper" or any individual who is not disseminating content as part of a "systematic business, professional or commercial activity" (section 1(t)). This would appear to cover any online news outlet as well as any individual who disseminated content professionally. Online curated content refers to any "curated catalogue of audio-visual content" which is made available on demand (section 1(q)).

The Brazilian Fake News Law would apply to "Internet applications" as defined in Article 5(VII) of the Marco Civil, namely as "a set of functionalities that can be accessed through a terminal connected to the Internet". We are not aware of how this may have been interpreted by Brazilian courts but it would appear to be very broad in nature. However, the Fake News Law is also limited to social media and private messaging services and, from among these actors, only to social media with at least two million registered users in Brazil (Article 1).

The Canadian Technical Paper covers "online communication service providers" (OSCPs), defined as providers of a "service that is accessible to persons in Canada, the primary purpose of which is to enable users of the service to communicate with other users of the service, over the internet", while specifically excluding private communications services (clause 2). It also envisages the possibility of regulations further clarifying the scope of the system (clause 3), which is currently not very clear and potentially very broad.

2.3. Content Covered

The DSA formally applies to all content which is illegal (or which may attract liability under the law), since it operates so as to remove liability as long as certain conditions are fulfilled. However, as highlighted in other parts of this Paper, it does a lot more than simply require intermediaries to take content down. Article 1(1)(3) of NetzDG specifically lists 20 different provisions of the German Criminal Code to which it applies. This covers a wide range of content crimes, including controversial ones from a human rights perspective such as criminal insult and defamation (both providing for imprisonment), blasphemy and merely glorifying violence, as well as more accepted ones such as child pornography and incitement to crime.

The Indian Rules define a wide range of content which is covered by the regime in section 3(1)(b). Formally, this takes the form of requiring these prohibitions to be included in the rules of the intermediary (and to inform users about those rules). The content covered includes anything that is contrary to any law (sections 3(1)(b)(ii) and (v)) or which is clearly illegal, such as information the dissemination of which infringes intellectual property rights (sections 3(1)(b)(i) and (iv)). It also covers some content which is likely partly legal and partly illegal, such as information which "threatens the unity, integrity, defence, security or sovereignty of India, friendly relations with foreign States … or is insulting other nation" (section 3(1)(b)(viii)) or which is disseminated with the intent to "mislead or harass a person, entity or agency for financial gain or to cause any injury to any person" (section 3(1)(b)(x)). Finally, it covers several categories of information which are presumably either entirely or largely legal in India, such as information which is harmful to children (section 3(1)(b)(iii)), which misleads the addressee about its origin (section 3(1)(b)(vi)), which is patently false or misleading (section 3(1)(b)(vi)), which impersonates another person (section 3(1)(b)(vii)), or which is "patently false and untrue".

The Indian Rules impose additional obligations on news publishers and on-demand services through a Code of Ethics (Appendix) which requires news publishers to respect the Norms of Journalistic Conduct of the Press Council of India, adopted under the Press Council Act, 1978, and the Programme Code adopted under section 5 of the Cable Television Networks (Regulation) Act, 1995 (Part I of the Code). Presumably many of the standards in these codes limit the dissemination of content which is otherwise legal. On-demand services are required to exercise "due caution and discretion" before featuring any content which "affects the sovereignty and integrity of India", "threatens, endangers or jeopardises the security of the State", "is detrimental to India's friendly relations with foreign countries" or "is likely to incite violence or disturb the maintenance of public order" (Part II(A) of the Code). Some but not all of this content would likely be illegal under Indian law.

The Brazilian Fake News Law prohibits Internet applications from allowing a number of types of behaviour on their services, namely "inauthentic accounts", "unlabelled artificial disseminators, understood as those whose use is not communicated to the application provider and the user as well as those used for dissemination of misinformation", and "artificial dissemination networks that disseminate misinformation", as well as "unlabelled sponsored content" (Article 5). It also requires measures to be taken against disinformation (Section III of Chapter II), defined as content which is "unequivocally and verifiably false or

misleading, taken out of context, manipulated or forged, with potential to cause individual or collective harm, with the exception of a humorous or parody spirit" (Article 4(II)). It may be noted that, outside of election contexts, Brazil does not have a general prohibition on disinformation or false news, so technically this is legal content (and this is further addressed under Legal Regulation of Legal Content below).

Like NetzDG, the Canadian Technical Paper is limited to certain kinds of essentially already illegal content, albeit its scope is much narrower, covering only five types of information, namely child sexual exploitation content, terrorist content, hate speech, content that incites violence and non-consensual sharing of intimate images. The definitions of this content, which are yet to be developed fully, should "borrow from the Criminal Code but are adapted to the regulatory context". While some limiting parameters are placed on these categories of information – such as that terrorism content should actively encourage terrorism and be likely to result in terrorism – the actual legislative proposal would need to clarify its scope far more precisely (clause 8). The proposal also requires OCSPs to "publish clear content-moderation guidelines, applicable to the five (5) types of harmful content", which guidelines may also be prescribed by regulation (clause 13).

2.4. Becoming Aware of Illegal Content

This part of the Paper focuses on what needs to happen for regulated entities to be deemed to have become aware of the illegal content (or, otherwise, what actions trigger their responsibilities under the rules). As noted above, it is neither legitimate nor practical to require intermediaries to monitor the vast number of communications that flow through their systems. The systems reviewed here do not, by and large, require them to do that and in some cases specifically rule that out. Thus, Article 7 of the DSA provides: "No general obligation to monitor the information which providers of intermediary services transmit or store, nor actively to seek facts or circumstances indicating illegal activity shall be imposed on those providers." Otherwise, hosting services are liable once they receive "actual knowledge" of illegal content, which they are deemed to have received once notice which meets certain conditions, set out below, is provided to them by anyone (Articles 5 and 14(3)).

The DSA also goes beyond simply providing for liability for hosting services once "actual knowledge" has been established. Pursuant to Article 14, these services are required to put in place mechanisms to facilitate the reporting of illegal content via notices in a "sufficiently precise and adequately substantiated" manner so that an intermediary could determine whether or not the content was illegal. The system should facilitate notices that include reasons why the content is considered to be illegal, the electronic location of the content, the name and email of the person filing the notice (except in certain cases of sexual offences and child pornography) and a statement of their good faith. Any notice which contains this material shall be deemed to give rise to actual knowledge. Intermediaries are required to confirm receipt of notices promptly, process notices in a "in a timely, diligent and objective manner" and indicate to the "complainant" the action taken and possibilities of redress, as well as if automated means were used to process the notice.

Apart from micro or small enterprises, online platforms are also, pursuant to Article 19 of the DSA, required to establish a system of trusted flaggers, working with the Digital Services Coordinator which each Member State is required to appoint. This status shall be awarded upon application where certain qualities have been shown to be present, such as that the applicant has expertise in identifying illegal content, is independent from any platform and carries out its activities in a "timely, diligent and objective manner". The European Commission shall keep a public register (database) of trusted flaggers. There is also a procedure for platforms to highlight that a trusted flagger has filed a significant number of substandard notices, as well as for removing flaggers who no longer meet the conditions (with a high number of mistakes presumably suggesting that this might be the case).

As noted above, formally the NetzDG revolves around the complaints mechanism, which triggers the obligation to take action. The mechanism needs to be "effective and transparent", as well as "easily recognisable, directly accessible and permanently available" to users. Platforms are required to take "immediate note" of complaints and check whether the content is illegal. The handling of complaints need to be monitored via monthly checks done by management, with deficiencies being "immediately rectified" and those responsible for processing complaints must receive training at least once every six months. The Federal Office of Justice may task an administrative agency with monitoring complaints procedures (various provisions in Article 1(3)).

The Indian Rules have more restrictive rules on when an intermediary is deemed to have actual knowledge of illegal content, which is when they are notified about it "by the Appropriate Government or its agency" (section 3(1)(d) of the Rules which is identical in this respect to section 79(3)(b) of the governing Act to which it refers).

The Brazilian Fake News Law is not very clear as to when the responsibility of Internet applications will be engaged. Article 5(§3) calls on them, taking into account the rapidly changing nature of inauthentic behaviour, to "develop procedures to improve protections of society against unlawful behaviour". Sanctions for breach of the law under Article 28 are applied only by the judiciary, so presumably the courts might develop some criteria for applying the different levels of sanctions. That article indicates that, in setting sanctions, courts should take into account various factors such as the seriousness of the breach (including the reasons for it and the individual and collective consequences of it), any recurrence of the activity and the economic capacity of the application (for fines). It also requires applications to provide accessible and visible mechanisms for users to report disinformation (Article 24(I)).

On the other hand, Article 21 of the Marco Civil is quite clear that applications that do not remove, in a diligent manner, content which displays "nudity or sexual activities of a private nature" after being informed about it by a user, shall be liable for any privacy breach relating to that material.

Here, again, the Canadian Technical Paper departs from the approach of other systems by requiring OCSPs to "take all reasonable measures, which can include the use of automated systems, to identify harmful content" within the five categories outlined above (clause 10).

What might be deemed to constitute "reasonable measures" is not yet clear. It also requires OCSPs to put in place "accessible and easy-to-use flagging mechanisms" for harmful content, along with an "accessible and easy-to-use opportunity to make representations, and compel an OCSP to promptly review and reconsider its decision" (clause 12). Presumably actual legislation would provide more details about how these systems should operate.

2.5. Measures Required to be Taken

This section of the Paper canvasses the different sorts of measures that regulated entities are required to take when they become aware of illegal content. The primary such obligation under the DSA, as noted above, is to act "expeditiously to remove or to disable access to the information" (Articles 4(1)(e) and 5(1)(b)). Article 8 also provides for specific orders by competent national judicial or administrative authorities to "act against a specific item of illegal content", while also placing some conditions on such orders (such as to include reasons why the information is illegal, the URL of the information and information about redress available to the intermediary as well as the user who provided the content). Intermediaries are also required to respond to orders to provide information about users, again as long as the orders meet certain conditions (Article 9). As noted above, hosting service providers must process complaints quickly (Article 14(6)) and provide detailed reasons to users whose content has been removed or rendered inaccessible (Article 15). Larger online platforms must "suspend, for a reasonable period of time and after having issued a prior warning" users who "frequently provide manifestly illegal content", as well as the processing of notices from complainants who "frequently submit notices or complaints that are manifestly unfounded", taking into account four factors which are listed, such as the number of instances of these forms of abusive behaviour (Article 20). These platforms must also promptly inform the law enforcement or judicial authorities of evidence suggesting the commission of a serious criminal offence involving "a threat to the life or safety of persons" (Article 21).

Under NetzDG, similarly, the primary action is to remove or block access to illegal content. This must be done within 24 hours for content "that is manifestly unlawful" and within 7 days for other content (albeit with some exceptions, including where content has been referred to a self-regulatory institution on which see below) (Articles 1(3)(2)(2) and (3)). Certain conditions accompany the removal of content such as that the content is retained as evidence and both the complainant and user are immediately notified of the removal and provided with reasons (Articles 1(3)(2)(4) and (5)).

The main measure under the Indian Rules is also for intermediaries to take down or block access to the content within 36 hours, once they have actual knowledge of it, failing which they lose their immunity in relation to that content (section 3(1)(d)). However, sexually explicit content, defined broadly to include "any material which exposes the private area", shall be taken down within 24 hours of a complaint (section 3(2)(b)). Intermediaries are also required to inform users that they may have their usage rights terminated for non-compliance with the intermediary's rules, regulations, privacy policy or user agreement (section 3(1)(c)).

It is not clear whether this is intended as an indirect requirement for intermediaries to provide for such termination in their user agreements but it would presumably require that. No conditions are set out for when termination might result from a breach of the terms and conditions.

For news publishers and on-demand services, the Authorised Officer (a senior official appointed by the responsible ministry) may also initiate a process leading to the interim emergency blocking of content "for which no delay is acceptable" (section 16 of the Rules) on grounds of protecting "sovereignty and integrity of India, defence of India, security of the State, friendly relations with foreign States or public order or for preventing incitement to the commission of any cognizable offence relating to above" (section 69A of the governing Act). Once put into place, such blocking must be brought before the Inter-Departmental Committee established to hear appeals within 48 hours and then the final decision to continue or remove the interim blocking shall be made by the Secretary of the Ministry of Information and Broadcasting (section 16 of the rules).

The Brazilian Fake News Law requires applications to take measures against disinformation, outlined below under Legal Regulation of Legal Content. Otherwise, for the prohibited behaviours, it seems to imply that these should be labelled, although this is not clear (Article 5(§2). As noted above, Article 21 of the Marco Civil is quite clear in relation to nude or sexual content that invades privacy, requiring it to be taken down.

Pursuant to the Canadian Technical Paper, OCSPs are required to act "expeditiously" to "address" all content that is flagged as harmful. Expeditious is defined as acting within 24 hours or such other period of time as may be prescribed, while address is defined as responding to the affected person or flagger indicating either that the content does not fall within the scope of harmful content or that it does and has been made inaccessible (clause 11). The proposals also call for OCSPs to report incidents of illegal behaviour to the law enforcement authorities, with two quite different options being offered for this at the moment (clause 20). They should also preserve relevant information about the harmful content so as to facilitate possible future law enforcement actions (clause 23).

2.6. Complaints Systems for Users

It is important, in the context of both automated and human-driven moderation measures against allegedly illegal content, to ensure that the user who posted the content has options to contest those measures. This allows for information that may not have been available in the original decision-making process to be presented, which may show that the posting was not in fact illegal, as well as to correct mistakes.

The DSA requires the intermediary to inform the user, at least by the time of removing or disabling access to content, of the fact and territorial scope of the measure, the facts relied upon in the decision-making process, whether automated means were employed, the legal or contractual ground relied upon (depending on the basis for the measure) and redress possibilities, including judicial redress (Article 15). Online platforms which are not micro or

small in size must also offer access to an "effective internal complaint-handling system", which is "easy to access, user-friendly and enable and facilitate the submission of sufficiently precise and adequately substantiated complaints" for at least six months following any decision to remove or disable access, or to suspend or terminate the service or the account. Where a complaint offers "sufficient grounds" to conclude that the original decision was in error, the measure applied should be reversed (Article 17).

The DSA also provides for an "out-of-court" dispute settlement system (Article 18). Bodies which wish to perform this function must apply to the Digital Services Coordinator of the relevant State for certification, which shall be granted where the body meets the requisite conditions, such as that it is independent of both users and intermediaries, has the necessary expertise, is capable of settling disputes in a "swift, efficient and cost-effective manner" and operates according to "clear and fair rules of procedure". Platforms must engage in good faith with this dispute settlement process and are bound by the decisions of the body. Where the body decides in favour of the user, the online platform shall reimburse fees and other reasonable expenses, while if it decides in favour of the platform, the user is not required to reimburse any expenses incurred by the platform. The DSA also establishes a right to complain about breaches of its rules to the Digital Services Coordinator.

The Indian Rules require all intermediaries to publish "prominently" on their websites the name and contact details of the Grievance Officer they are required to appoint, as well as how to make a complaint about "violation of the provisions of this rule or any other matters". Complaints shall be acknowledged within 24 hours and disposed of within 15 days (sections 3(2) and (3)). Significant social media intermediaries shall in addition enable complainants to track the status of their complaints and provide reasons for any action taken or not taken (section 4(6)). They shall also inform users prior to removing or disabling access to any content they are responsible for of their intention to do so and the reasons therefor, as well as provide the user with an "adequate and reasonable opportunity" to dispute the action (section 4(8)).

The Rules also put in place a complex, multitiered complaints system for news publishers and on-demand services, described in more detail under Engaging Regulatory, Co-regulatory and Self-regulatory Bodies. For now, it may be noted that the framework rules for these complaints require an acknowledgement to be provided within 24 hours and the complaint to be addressed within 15 days (sections 10(2) and (3)).

Article 11 of the Brazilian Fake News Law, in the section on disinformation, requires applications to notify the user who first disseminated content, as well as anyone else who shared it, about any measures taken against such content, the reasons for taking those measures and the sources used in verifying (presumably the inaccuracy of the information). Article 12 requires applications to provide an accessible and visible complaints system, available for at least three months after a decision against content has been taken, so that the user who first disseminated the content, as well as the author, if different, can lodge an appeal and present information as to the veracity of the information. Where proof of veracity is accepted, the application should reverse its original decision.

As noted above, the Canadian Technical Paper envisages both complaints (or flagging) to OCSPs and also the possibility of an internal appeal to compel the OCSP to review its decision "promptly". Notice of both the original measure taken and of its "reconsideration", along with information about the right to appeal further to the Digital Recourse Council of Canada (Council) should be provided by OCSPs, presumably to both the original flagger and the affected user (although this is not spelt out very clearly) (clause 12). The Council may receive written complaints about both a failure to render content inaccessible and a decision to render it inaccessible, but only after the complainant has "gone through the relevant content moderation and reconsideration processes at the OCSP level" (clauses 49 and 50). Where the Council determines that content which has been maintained as accessible is harmful content, it shall issue a binding order to the OCSP to make that content inaccessible (clause 55).

2.7. Institutional Measures

Many systems envisage the putting in place of quite complex institutional structures to accompany the rules they establish. It is beyond the scope of this paper to get into too much detail about these institutional systems. However, in some cases they affect the way these systems reflect and respect rights so to that extent they are reviewed here.

The DSA puts in place complex institutional arrangements. Importantly, intermediaries which do not have an establishment in the EU must designate a legal or natural person as their legal representative, who can be held liable for non-compliance, and allocate that person "the necessary powers and resource to cooperate" with various other engaged institutions (Article 11). As such, intermediaries cannot seek to avoid liability simply by not being established within the EU (there is also a provision on "traceability of traders" in Article 22). Very large platforms also have to appoint compliance officers who are generally responsible for monitoring their compliance with the DSA rules and cooperating with the Digital Services Coordinator, organising the independent compliance audit and advising management and employees (Article 32).

Under the Indian Rules, significant social media intermediaries must appoint a Chief Compliance Officer who is responsible for ensuring compliance with the Act and Rules and who shall bear directly liability for any third-party content where he or she fails to ensure that the intermediary exercises due diligence in discharging its obligations (section 4(1)(a)). They much also appoint a "nodal contact person" who is available "24x7" for purposes of coordinating with law enforcement agencies (section 4(1)(b)). As noted above, they must also appoint a Grievance Officer who must be resident in India and perform the duties described above under Complaints Systems for Users (section 4(1)(c)), and have a "physical contact address in India published on its website" (section 4(5)).

The Brazilian Fake News Law requires both social media platforms and private messaging services to appoint legal representatives in Brazil and provide their information in an accessible manner on their digital platforms (Article 29). More generally, the Marco Civil requires all levels of government to follow various guidelines in the development of the

Internet, including by having "mechanisms of governance that are multi-stakeholder, transparent, cooperative and democratic" (Article 24(I)).

The Canadian Technical Paper envisages a rather complicated set of institutions, including a Digital Safety Commission, a Digital Safety Commissioner, a Digital Recourse Council and an Advisory Board. Broadly speaking, the Commissioner exercises general oversight powers, including to order OCSPs to take steps to comply with the rules, while the Council is the appellate body for decisions by OCSPs in relation to specific content. Interestingly, the Commissioner, with the approval of Treasury Board,[55] may adopt regulations setting out the "regulatory charges that one or more classes of OCSPs must pay … to recover the costs of the Commission, the Digital Safety Commissioner, and the Digital Recourse Council of Canada" (clause 66). In other words, regulated entities are to pay for the operation of the system.

2.8. Other Issues

The Indian Rules also require significant social media intermediaries to "endeavour to deploy technology-based measures" to proactively identify and remove "information that depicts any act or simulation in any form depicting rape, child sexual abuse or conduct", as well as information which is identical to other information which has already been removed. Users attempting to access such content are to receive a notice that it has been banned. This system must be "proportionate", taking into account the "interests of free speech and expression, privacy of users", including by being subject to "appropriate human oversight" (section 4(4)).

The Indian Rules also require on-demand services to classify all content according to age appropriate standards broken down into "U" (open), "U/A 7+" (suitable for those aged 7 and above), "U/A 13+" (suitable for those aged 13 and above), "U/A 16+" (suitable for those aged 16 and above), and "A" (for adults) (Code, Part II(B), (C), (D) and (E) and Schedule).

The Brazilian Fake News Law requires private messaging services to put in place a number of structural measures to limit or slow down the spread of disinformation. For example, Article 13 requires a limit of five to be placed on the forwarding of a message to users or groups and groups to be limited to 256 members. During electoral periods, this is limited to one forward. According to Article 15, the user must authorise the use of mass communication features such as broadcast lists or group chats.

In South Africa, the Films and Publications Amendment Act, 2019[56] came into effect on 1 March 2022. It required Internet service providers (ISPs), defined as entities which provide access to the Internet, to register with the Films and Publications Board and to take measures, among other things, to protect children from exposure to pornography and to limit the spread

[55] The role of this body is defined by the Government of Canada as being "responsible for accountability and ethics, financial, personnel and administrative management, comptrollership, approving regulations and most Orders-in-Council". See https://www.canada.ca/en/treasury-board-secretariat/corporate/about-treasury-board.html.
[56] Act No. 11 of 2019, 3 October 2019, https://www.ellipsis.co.za/wp-content/uploads/2019/07/Films-and-Publications-Amendment-Act-11-of-2019.pdf.

of "child sexual abuse material". In a Notice issued on 28 October 2022,[57] the Films and Publications Board required registered ISPs to report to it on how they: had taken reasonable steps to ensure that any "child-oriented" services (services aimed at children) were not being used to commit an offence against children; had displayed safety messages on all advertisements on child-oriented services and in chatrooms or similar "contact services" (places where people meet); had provided a mechanism to enable "children to report suspicious behaviour by any person in a chatroom"; had reported behaviour suggesting the commission of an offence to the police; and, where technically feasible, had provided children and their parents with information about software and other tools to block access to services where access by a child would constitute an offence under the Act (clause 2).

In addition, where an ISP had knowledge that its services were being used to disseminate "child pornography, propaganda for war, incitement of imminent violence or advocating hatred based on an identifiable group characteristic and that constitutes incitement to cause harm", it was required to report on the reasonable steps it had taken to prevent access to this content, whether they had reported the offence to the police, and any reasonable steps they had taken to preserve evidence of the crime (clause 5).

Finally, all online service providers who were aware that their service was being used to host or distribute "unclassified content, prohibited content, or potential prohibited content" should report on referrals of that content to the Films and Publications Board, reasonable steps taken to prevent the use of their service to host or distribute that content, any takedown notices they had issued and the content which they had taken down (clause 8).

In some cases, laws attempt to place wider positive obligations on the public as a whole. For example, in its Organic Law for the comprehensive protection of children and adolescents against violence,[58] Spain requires any person, physical or legal, "who notices the existence of content available on the Internet that constitutes a form of violence against any child or adolescent" to report that to the competent authority and, if the facts seem to represent a crime, to the law enforcement authorities. For their part, "public administrations" must guarantee the availability of accessible and secure channels for reporting on such content (Article 19).

3. Legal Regulation of Legal Content

This section of the Paper focuses on legal (i.e. mandatory) measures which address speech which is otherwise legal (i.e. for the author to disseminate) but arguably harmful ("lawful but awful"), with a focus on systemic measures set out in law to address this speech. By definition this is more controversial than the previous section since it provides for measures to address content which is otherwise legal, often including taking down such content. The

[57] Notice No. 2682, 28 October 2022, https://www.ellipsis.co.za/wp-content/uploads/2022/11/FPB-Notice-Gazette-47373-28-10.pdf.
[58] Law 8/2021 of 4 June, available in Spanish at: https://www.boe.es/buscar/act.php?lang=en&id=BOE-A-2021-9347&tn=1&p=.

core idea behind these sorts of legislative rules is that, under certain conditions, taken collectively and often due to its volume, this content may create harms via platforms even though the individual speech acts covered do not rise to the level where they could legitimately be sanctioned. In many cases, these rules seek to prevent direct harm to users, with a particular focus on children but often also adults, but another justification for this sort of measure is that this speech may harm the rights to "seek and receive" information, for example if at some point the volume of mis-/disinformation undermines users' right to seek information and, indeed, their ability to discern the truth.

The Australian Online Safety Act 2021, which has been in effect for about a year, is perhaps the most far-reaching effort to regulate this sort of content and so it occupies a particular focus in this section, which outlines in some detail its different provisions, broken down into the types of content covered, the systemic approaches to addressing this content (such as notice and response) and the types of measures taken in relation to this content. This section then reviews briefly the extent to which a number of other systems, most of which are reviewed in more detail in the previous section, also address otherwise legal speech.

3.1. The Australian Online Safety Act 2021

The Australian Online Safety Act 2021 came into effect only about a year ago on 23 January 2022.[59] It defines a range of actors to which its provisions are variously applicable, including social media services (section 13), relevant electronic services (section 13A, mostly direct communications services like email and SMS), designated Internet services (section 14, services which allow users to access material), hosting services (section 17) and on-demand programme services (section 18).

Key to the functioning of the Australian system is the eSafety Commissioner, created by the Act, with very broad functions and powers (described in general terms in sections 27 and 28 but elaborated on in many other sections). The Commissioner is appointed by the minister who is responsible for the Act, apparently in his or her discretion although subject to limited conditions of expertise (section 167). One of the criticisms of the Act is the very broad and often unfettered role and powers allocated to the Commissioner.[60]

- Types of Content

The Act could be said to establish seven different content and practice regimes relating, respectively, to cyber-bulling (of children), intimate images, cyber-abuse (of adults), "abhorrent violent conduct" material, "basic online safety expectations", the "online content scheme", and industry codes and standards. Of these, only "abhorrent violent conduct" material, for which the Act addresses depicting, promoting, inciting and instructing in (that

[59] Available at: https://www.legislation.gov.au/Details/C2022C00052/Download.
[60] See Malcolm Campbell, *Australia: Tightening the law around online content: Introduction of the Online Safety Act 2021*, 31 May 2022, https://www.mondaq.com/australia/security/1197494/tightening-the-law-around-online-content-introduction-of-the-online-safety-act-2021-cth.

sort of content), is clearly illegal pursuant to the Criminal Code (Subdivision H of Division 474).[61]

Cyber-bullying is defined as material which a reasonable person would conclude was intended and likely to have a "seriously threatening, seriously intimidating, seriously harassing or seriously humiliating" effect on a particular Australian child (someone who is under 18 years old) (section 6(1)). Some of this material might be illegal in Australia but much of it would not be.

Intimate images covers both "depiction of private parts" – namely depiction of the genital or anal areas or the breasts of a "female person or transgender or intersex person" – in circumstances in which a reasonable person would expect to be afforded privacy and "depiction of private activity" – covering a range of activities such as undressing, showering or engaging in a sexual act – again in circumstances in which a reasonable person would expect to be afforded privacy (section 15). Important exceptions to this are provided in sections 75(3) and (4) and 86, for example relating to law enforcement or judicial proceedings.

Cyber-abuse (targeting adults) is defined as material which a person in the position of the targeted adult would regard as being, "in all the circumstances, menacing, harassing or offensive" and which a reasonable person would conclude was intended to cause "serious harm to a particular Australian adult" (section 7(1)). In determining whether material is offensive, the following should be taken into account: "the standards of morality, decency and propriety generally accepted by reasonable adults"; "the literary, artistic or educational merit (if any) of the material"; and "the general character of the material (including whether it is of a medical, legal or scientific character)", as well as whether consent was given to its publication (section 8). Again, much of this material would not only be legal but is merely offensive, a type of content that international courts have repeatedly stressed is protected by the right to freedom of expression.

The online content scheme is essentially defined in reference to the classification standards under the Classification (Publications, Films and Computer Games) Act 1995. Class 1 content is defined as content which has either been classified as RC (refused classification) or would be so classified if it had been classified (for material which has not been assessed for classification) (section 106). According to the Australian Government:

> Refused Classification (RC) is a classification category referring to films, computer games and publications that cannot be sold, hired, advertised or legally imported in Australia. RC-classified material contains content that is very high in impact and falls outside generally-accepted community standards.[62]

Class 2 covers two types of content, that which has either been classified as Restricted (X 18+) or would be so classified if it had been classified and content which is similarly covered by

[61] As such, the rules relating to that content are not elaborated upon here.
[62] "What do the ratings mean?", https://www.classification.gov.au/classification-ratings/what-do-ratings-mean.

Restricted (R 18+) classification (section 107). Regarding the former, and again according to the Australian Government:

> X 18+ films are restricted to adults. This classification is a special and legally-restricted category due to sexually explicit content including actual sexual intercourse or other sexual activity between consenting adults. X 18+ films are only available for sale or hire in the ACT [Australian Capital Territory] and some parts of the NT [Northern Territory].[63]

R 18+ material is high-impact material which is limited to adults and may be considered to be offensive by some adults. R X 18+ material should generally not be disseminated (and may be subject to a removal notice if it is) while R 18+ material should either not be disseminated or be subject to a restricted access system which effectively prevents children from accessing it. Once again, while some of this material, presumably mostly only Class 1 material, though, would be illegal, much of it would not be.

The other two regimes – namely "basic online safety expectations" and industry codes and standards – relate more to expected behaviour on the part of regulated services.

- Systemic Approaches

The Act provides broadly for complaints to be made to the Commissioner about breach of the rules relating to cyber-bullying (section 30), intimate images (sections 32 and 33), cyber-abuse (section 36) and the online content scheme rules (section 38). The Act also provides for complaints to be made to the Commissioner about breach by service providers of their own rules (terms and conditions) or industry codes (sections 39 and 40). Where a complainant is seeking a removal notice in relation to cyber-bullying or cyber-abuse content, they must show that they made a complaint to the service provider which did not result in the content being taken down before complaining to the Commissioner. The different measures that may be taken in response to complaints are detailed below, under Types of Measures, but broadly speaking takedown of the content is a primary response.

The basic online safety expectations system is basically driven by the Minister, via regulation, which can set out the content of these expectations for different types of service providers, referred to as determinations. Section 46 of the Act sets out nine types of expectations which must be included in each determination, namely that: the provider will take reasonable steps to ensure that the service is safe for users (1); that in doing so the provider will consult the Commissioner (2); that the provider will take reasonable steps to minimise the presence of the various types of content described above on their services (and to ensure that "technological or other measures are in place" to prevent access to Class 2 content by children) (3 and 4); that providers have in place "clear and readily identifiable mechanisms" for users to lodge complaints about content addressed by the Act or breaches of the provider's own terms of use (5 and 6); and that providers will comply with requests by the Commissioner for reports to him or her on the number of complaints made, on how long it

[63] *Ibid.*

took the provider to comply with removal notices and on measures taken to ensure that users are able to use the service safely (7-9).

Finally, the Act envisages a system of industry codes and standards. It sets out Parliament's expectation that bodies representing "sections of the online industry" should develop industry codes for their part of industry and that the Commissioner should either register a code for each section of the industry or adopt its own industry standards for that section (section 137). Section 138 sets out a long list of examples of matters that may be dealt with by codes and standards which focuses on issues like procedures for addressing various matters, with a focus on class 1 and 2 material, promoting awareness and providing information to users and parents. Where the Commissioner is satisfied that a code provides appropriate community safeguards and has been the subject of adequate consultation it may register that code (section 140). The Commissioner may also request representative bodies to adopt codes dealing with specified matters (section 141). And where the Commissioner is satisfied that a code has been breached, it may direct the relevant party to comply with that code (section 143). Where either a requested code has not been developed or does not meet the conditions for registration, or a registered code has become outdated and is not being updated, and it is "necessary or convenient" to provide for appropriate community standards, the Commissioner may adopt an industry standard governing the relevant matters (section 145). A failure to comply with such a standard may be subject to a civil penalty and a formal warning (sections 146 and 47).

The Act also provides for the Commissioner, by regulation, to adopt binding rules pursuant to "service provider determinations". These are essentially limited to what is authorised by paragraph 51(v) of the Constitution (section 151), which empowers the federal government to adopt rules relevant to peace, order and good government in the areas of "postal, telegraphic, telephonic, and other like services".[64] Breach of the rules in these determinations may also lead to a civil penalty (section 153).

- Types of Measures

As noted above, removal notices are a primary possible response to much of the material that is addressed by the Act. Thus, such notices, normally to be acted on within 24 hours, may be issued to both service providers and the person responsible for posting the material for cyber-bullying material (sections 65 and 66, if it was not taken down by the provider within 48 hours, and section 70 for the user), intimate images (section 77 and section 78 for the user), cyber-abuse (sections 88 and 90, if it was not taken down by the provider within 48 hours, and section 89 for the user), for class 1 material (sections 109 and 110), and class 2 (X 18+) material (sections 114 and 115). For class 2 (R 18+) material, the provider has the option either of removing the material or restricting access to the material to adults (sections 119 and 120). In most cases, failure to comply with a removal notice may lead to a civil penalty.

[64] Section 51 of the Commonwealth of Australia Constitution Act is available at: https://www.aph.gov.au/About_Parliament/Senate/Powers_practice_n_procedures/Constitution/chapter1/Part_V_-_Powers_of_the_Parliament#chapter-01_part-05_51.

For cyber-bullying material, the Commissioner may also issue a notice to the user who is responsible for posting the material to refrain from posting addition cyber-bulling material targeting that child and to apologise to the child (section 70). For intimate images, the person responsible for posting the material may also be subject to a civil penalty (section 78).

For material depicting, promoting and so on abhorrent violent conduct, the Commissioner may issue a notice to the service provider to block access to the material (sections 95 and 99). And for class 1 material, search engines may be required to cease to provide links to the material (section 124), while app distribution services may, in relation to apps that facilitate "the posting of class 1 material" be required to prevent users in Australia from downloading that app (section 128).

Where the Commissioner believes that a provider has breached a basic online safety expectation, it may make a statement to that effect and send the statement to the provider and publish it (section 48). More systemically, the Commissioner may require periodic reporting by a provider (section 49) or class of providers (section 52) on one or more basic online safety expectations, or non-periodic reporting again by a provider (section 56) or class of providers (section 59) on one or more basic online safety expectations. Failure to produce such a report may lead to a civil penalty and a warning (respectively, sections 50, 51, 53, 54, 57, 58, 60 and 61).

As noted above, breach of an industry code may result in a direction to comply with that code (section 143), while a failure to comply with an industry standard may lead to a civil penalty and a formal warning (sections 146 and 47). Breach of a rule adopted pursuant to a "service provider determination" may lead to a civil penalty (section 153).

Importantly, where the Commissioner is satisfied that a provider has, twice within the previous 12 months, contravened a civil penalty provision of the online content scheme, he or she may apply to the courts for an order terminating the service (section 156).

3.2. Other Approaches

The primary focus of the DSA is on illegal (liable) content. However, it does provide some support for self-regulatory measures, addressed in the next section, as well as transparency in this space, addressed in the section after that. Here, we will just mention Article 12(2), which requires intermediaries, in relation to any restrictions they impose through their terms and conditions in respect of information provided by users, to "act in a diligent, objective and proportionate manner in applying and enforcing" such restrictions. This is significant since it effectively requires intermediaries to apply their content rules properly.

As noted above under Content Covered, the Indian Rules, while focusing primarily on illegal content, also cover quite a range of content which is presumably legal under Indian law. Examples of this include the requirement for intermediaries to ban, via their own rules, content that is harmful to children, misleads about its origin, is patently false or misleading

or impersonates another person (section 3(1)(b)).[65] Such information, once properly identified, is required to be taken down. News publishers must also follow the Norms of Journalistic Conduct of the Press Council of India and the Programme Code under the Cable Television Networks (Regulation) Act (Code of Ethics, Part I), both of which presumably go beyond addressing illegal content.

The Brazilian Fake News Law generally requires applications to take "necessary measures" to protect society against disinformation, which is not otherwise generally illegal under Brazilian law. Such measures should be proportionate, not discriminatory and not imply a restriction on the "free development of the personality individual, artistic, intellectual, satirical, religious, fictional, literary or any other form of cultural manifestation" (Article 9). The Law does not set out what specific measures should be used, but Article 10 outlines a number of possible good practices. These include:

- the use of verifications from independent fact-checkers;
- disabling the capacity to forward disinformation to more than one user at a time;
- labelling disinformation as such;
- stopping, rapidly, the paid promotion or "artificial free promotion" of the disinformation; and/or
- sending verified information to all users who accessed the disinformation.

This is an interesting approach inasmuch as it refers to a wider range of possible measures, including measures against recommender systems (as part of the fourth bullet above), to address disinformation.

The United Kingdom Online Safety Bill underwent a significant transformation in terms of the scope of legal content that it addresses. In the June 2022 version,[66] sections 12 and 13 of the Bill required certain regulated services (covering larger both social media and related platforms and search engines) to conduct specific "adult risk assessments" and then to perform a number of duties to protect adults, including by setting out in the terms of service the manner in which it would deal with "priority content that is harmful to adults" such as by taking it down, restricting access to it or limiting the promotion of it (sections 13(3) and (4)). What constituted "priority content that is harmful to adults" was to be designated in regulations made by the Secretary of State (section 54(2)). But content that is "harmful to adults" was defined as content which presents a "material risk of significant harm to an appreciable number of adults", along with priority content which is harmful to adults

[65] India does not have a law generally banning false news although it does have some provisions prohibiting certain types of false statements. See Khushbu Jain & Brijesh Singh, "View: Disinformation in times of a pandemic, and the laws around it", 3 April 2020, The Economic Times, https://economictimes.indiatimes.com/news/politics-and-nation/view-disinformation-in-times-of-a-pandemic-and-the-laws-around-it/articleshow/74960629.cms#:~:text=Even%20though%20India%20does%20not%20have%20a%20specific,weapon%20that%20affects%20the%20morale%20of%20the%20people.

[66] Which was published after the Bill had gone through the Public Bill Committee, https://publications.parliament.uk/pa/bills/cbill/58-03/0121/220121.pdf.

(section 54(3)). These provisions attracted a lot of criticism and debate in the United Kingdom with the result that they have been removed from the latest (January 2023) version of the Bill.[67]

The Bill still has provisions on protecting children against harmful (but otherwise legal) material, including requirements generally to put in place proportionate measures to mitigate and manage the risks of harm to children in different age groups, as well as to prevent children from accessing "primary priority content that is harmful to children" and "other content that is harmful to children", including by measures such as blocking access of users to the service or particular content, or content moderation (such as takedowns) (sections 11(2), (3) and (5)). The Secretary of State shall, by regulation, designate "primary priority content that is harmful to children" and "priority content that is harmful to children", while harmful content also includes, in addition to these two categories, content which "presents a material risk of significant harm to an appreciable number of children in the United Kingdom" (sections 54(2), (3) and (4)). Harm includes physical or psychological harm (section 205). However, as Global Partners Digital noted in its September 2022 Digest, this effectively requires regulated services to choose between two "undesirable options", namely banning all material which may be harmful to children of any age or putting in place reliable age verification systems so as to block underage users.[68]

4. Engaging Regulatory, Co-regulatory and Self-regulatory Bodies

Most of the systems being put in place envisage some role for an official administrative regulatory body, whether an existing telcoms/broadcast regulator, election regulator or a newly created body with a specific mandate to regulate platforms or other digital actors. A wide range of different roles are envisaged for these bodies, some of which have been outlined in the previous two sections. Some of the powers of these regulators have also been highlighted in the previous two sections. Most of the laws reviewed impose broad obligations on regulated entities to cooperate with regulators, including by providing them with a wide range of information as relevant to regulatory objectives and by participating in hearings or reviews.

It is well established under international law that bodies which exercise regulatory powers over the media should be independent of both government and the sector they regulate.[69]

[67] The Bill went back to and was approved by the Public Bill Committee following those and other amendments and was then passed in a third reading by the House of Commons and is now before the House of Lords. See: https://bills.parliament.uk/bills/3137.
[68] Global Partners Digital: The Digest, September 2022, https://us11.campaign-archive.com/?u=3f454c6af66e369bd02cf4ac4&id=fa6f67f955.
[69] See, for example, Centre for Law and Democracy and International Media Support (IMS), Briefing Note Series on Freedom of Expression: Briefing Note 4: Independent Regulation of the Media, 2015, http://www.law-democracy.org/live/wp-content/uploads/2015/02/foe-briefingnotes-4.pdf.

The reasons for this are fairly obvious. As the Centre for Law and Democracy and International Media Support stated in a 2015 Briefing Note:

> In many countries, political interference in regulatory bodies has historically been the main concern but, in others, the greater threat is of regulatory capture by powerful commercial media players. Regulators which are properly insulated against both political and commercial influences are best able to perform their duties in the public interest.[70]

There are numerous authoritative international standards supporting this idea. For example, in its 2011 General Comment No. 34 on Article 19 of the International Covenant on Civil and Political Rights (ICCPR) the UN Human Rights Committee stated:

> It is recommended that States parties that have not already done so should establish an independent and public broadcasting licensing authority, with the power to examine broadcasting applications and to grant licenses.[71]

While this statement focused on broadcasting, other statements cover the media more broadly. For example, in their 2003 Joint Declaration, the (then) three special international mandates on freedom of expression highlighted the need for independence of all media regulatory bodies:

> All public authorities which exercise formal regulatory powers over the media should be protected against interference, particularly of a political or economic nature, including by an appointments process for members which is transparent, allows for public input and is not controlled by any particular political party.[72]

General Comment No. 34 focused on broadcasting, because many States do not have bodies which regulate the print media. Similarly, these earlier statements did not envisage specialised regulators for digital communications and so focused on the media. But the rationale for independence applies just as much to bodies which regulate digital communications as to media regulators. And many of the systems described above incorporate clear and strong rules on independence from both government and platforms.

One of the questions which almost immediately arises in this context is how to deal with the enormous volume of potential complaints and oversight issues that might flow to a regulator. This imposes both financial costs, which must be paid for from the public purse, and human resource costs, given that expertise is needed to address complaints and other oversight issues properly. A closely related issue is technical expertise relating to the way platforms actually function. Given that many of the automated systems that largely run most platforms are covered by commercial secrecy, there may be limits to how far regulators can probe into their functioning.

Given the recent vintage of most of these systems – the DSA only came into effect on 16 November 2022 and the Australian Online Safety Act 2021 only about a year ago, while the British law is still being developed – it is not yet possible to assess reliably how the volume issue will play out in practice. But one solution to it which is promoted by some of these

[70] *Ibid.*
[71] Note 5, para. 39.
[72] Adopted 18 December 2003, https://www.osce.org/fom/28235.

systems, such as the DSA and Indian Rules, but not by others, such as the Australian Online Safety Act, is to use a tiered system for complaints. There are different models for this but a key feature of most is that complaints initially go to the platform to resolve (while many platforms already have such systems in place of their own initiative for their own terms and conditions of service). This would presumably lead to a final resolution of a large majority, perhaps the vast majority, of complaints.

In many cases, the rules then provide for an appeal to an independent but not statutory body, which could be described as a co-regulatory level of appeal. Thus, the DSA provides for an "out-of-court" dispute settlement system, which will presumably mostly be run by private, for-profit bodies although different options may also emerge, while the Indian Rules provide for a "self-regulatory" body, presumably to be developed by sector players but ultimately approved by the Minister (see below). Both systems put in place conditions for recognising actors as being competent to operate at this level of appeal. For the DSA, the body must be independent (of both users and intermediaries), have the requisite expertise, and be able to settle disputes quickly and cheaply and operate fairly. The Indian Rules require the body to be able to oversee adherence to the rules, decide complaints within 15 days, provide advice to those covered by the rules and be able to impose a range of listed remedies.

Both the DSA and the Indian Rules also envisage final resolution of disputes by a statutory body, the Digital Services Coordinator in the case of the DSA and the Inter-Departmental Committee in the case of the Indian Rules.

The recent implementation of many of these systems means that new approaches to managing regulatory responsibilities among different types of regulators – i.e. statutory, and co- and self-regulatory bodies – have not yet demonstrated their strengths and weaknesses. In other words, it is still too early to assess this. But one feature we are seeing being built into these systems is an attempt to ensure flexibility to respond to the rapid technological and innovative changes to digital communications systems that we are currently witnessing. Thus, the section on Impact or Due Diligence of this Paper describes obligations in some systems for regulated entities to conduct risk assessments which look at the negative impacts of platform operations in a number of areas – such as illegal content, or on human rights or children – and then to put in place mitigation measures to address these. While this is perhaps not a self- or co-regulatory approach in the traditional sense of that term, it can certainly be seen in this light. Much of the harm that flows from platform activity is generalised in nature rather than about a failure to address a specific instance of harmful content and so regulation must also focus on that level, as impact assessments do. A similarly collaborative approach is seen in the Australian Online Safety Act, which calls on sector actors to develop industry codes in different areas, including as prompted by the regulator, the Commissioner, but also allows the Commissioner to step in and effectively impose industry standards where the codes have not been developed, are not fit for purpose or have become outdated and not been updated.

Also falling within the scope of coregulation are rules, increasingly found in the new laws, requiring regulated entities to enforce their own terms and conditions fairly and effectively.

Thus, as noted above, Article 12 of the DSA requires intermediaries to "act in a diligent, objective and proportionate manner in applying and enforcing" their own terms of service. Coupled to this is the approach taken in many of the new rules which provide for enforcement by requiring regulated entities to incorporate the standards the rules prescribe into their own terms and conditions. While this is not, itself, a co-regulatory approach, the fact that enforcement of those terms and conditions starts with the regulated entity itself means that it does become part of the system of co-regulation.

In terms of specific measures in this area, in addition to some of the more general approaches outlined above, while the primary focus of the DSA is on illegal content, not on self- or even co-regulatory measures, it does include a couple of "soft" measures to support non-binding measures. Article 34 provides generally that the EU Commission shall "support and promote the development and implementation of voluntary industry standards" in a number of areas, such as the submission of notices and auditing (of very large platforms). More specifically on this point, the Commission and Board (essentially composed of the Digital Services Coordinators) shall "encourage and facilitate the drawing up of codes of conduct" at the level of the EU both to contribute to the implementation of the DSA and on advertising (Articles 35 and 36).

NetzDG has wired into its rules quite a significant potential role for self-regulatory initiatives. The time limit for the primary obligation to remove unlawful content within seven days may be exceeded if the platform refers the decision on unlawfulness to "a recognised self-regulation institution" and agrees to accept the decision of that institution (Article 1(3)(2)(3)(b)). A body shall be recognised as a self-regulation institution if certain conditions are met, including that its independence and expertise are ensured, it can guarantee an analysis of content within seven days, it has set rules of procedure which provide for the possibility to review decisions, a complaints system has been set up and it is funded by several platforms, and it is open to others, "guaranteeing that the appropriate facilities are in place" (Article 1(3)(2)(6)). The Federal Office of Justice is empowered to recognise bodies as self-regulation institutions (Article 1(3)(2)(7)). At least one body has been recognised as a self-regulation institution under NetzDG.[73]

The Indian Rules set out quite an expansive system of self regulation for news publishers and on-demand services (both referred to in the Rules as "publishers"). To some extent this is different from platforms, as these entities are more like media outlets than digital intermediaries. But an overview of the system is provided here for completeness. The system requires three levels of appeal to be put into place, first internally within the publisher, second to a sector-wide body and third to an official oversight mechanism. For the first, each publisher must appoint a Grievance Officer based in India who is responsible for dealing

[73] The Voluntary Self-Regulation for Multimedia Service Providers (FSM) was apparently the first body to be recognised by the Federal Office of Justice as a self-regulation institution under NetzDG. See https://www.inhope.org/EN/articles/fsms-role-of-self-regulation-under-the-german-network-enforcement-act.

with complaints based on the Code of Ethics (which is set out in the Appendix to the Rules) and publish his or her name and contacts on its website (sections 11(1)-(3)).

Publishers must also belong to a "self-regulatory" body headed by a retired judge or other independent, eminent person and having up to six other expert members, and which is registered with the Ministry upon it being satisfied that the body meets the conditions set out in the Rules and is ready to process appeals from the internal complaints system. This body can warn the publisher or advise it to publish an apology, reclassify on-demand content or, in certain cases, advise the deletion of the material. Where the publisher fails to take action in accordance with such advice, the self-regulatory body may refer it to the Oversight Mechanism, or third level of complaint (section 12).

The key body within the Oversight Mechanism is the Inter-Departmental Committee, appointed by the Ministry of Information and Broadcasting and comprising representatives from the Ministries of Information and Broadcasting, Women and Child Development, Law and Justice, Home Affairs, Electronics and Information Technology, External Affairs, Defence, and such other ministries, along with "domain experts", as the Ministry of Information sees fit. The Authorised Officer, a senior officer from the Ministry of Information and Broadcasting, shall chair the Committee. The Committee is responsible for hearing appeals from self-regulatory bodies and may make similar recommendations to the Minister as the advisories by self-regulatory bodies. The Minister may, considering the recommendations of the Committee, issue orders to publishers to act (section 14).

An interesting initiative in South Africa is the Digital Complaints Committee (DCC), an essentially civil society initiative run by the non-governmental organisation Media Monitoring Africa under the banner "Real 411: Fight Disinformation Together",[74] but which has gained formal recognition from official bodies such as the Electoral Commission of South Africa.[75] It focuses on four areas, namely harmful false information, hate speech, incitement to violence and harassment, and provides an online system for individuals to report instances of these types of content. The complaint is then assessed by a sub-committee and referred to an appropriate third party, such as a self-, co- or statutory regulatory body, a platform or the courts, or other action may be taken, such as a counter-narrative published.[76] Although the number of complaints is relatively low – on average only one or less per day with an all-time high of 19 on 1 June 2022 – this represents an interesting model of informal cooperation between a civil society initiative and official and commercial actors.

5. Indirect Measures to Support Content Regulation

This section reviews other legal obligations imposed on different regulated entities to take actions which serve as indirect supports for the content measures outlined above. An important focus here is on transparency requirements in different areas. This, in turn, is

[74] See https://www.real411.co.za.
[75] See https://elections.real411.org.za/about.
[76] See https://www.real411.co.za/complaints-process.

broken down into transparency obligation regarding content moderation standards and systems, including automated systems (algorithms) that affect content, a special section on recommender systems, regular broader reporting obligations, many of which cover individual and State-driven content moderation requests and responses to them, as well as self-action by regulated entities, and special rules around transparency of advertising systems and how they work.

A second focus area here is on requirements to conduct (and publish) impact and/or due diligence assessments or audits, whether focusing on human rights, different specific areas of risk or other issues. This also covers requirements to take action in response to such assessments or audits, such as mitigation measures, reporting and/or engaging oversight systems.

Finally, this section provides a brief overview of legal requirements in the area of media and information literacy. This includes both general obligations, often on oversight bodies, to engage in awareness raising or literacy efforts, as well as sectoral obligations in this area. Given the broad and disbursed nature of these sorts of obligations, only a few illustrative examples are provided here.

5.1. Transparency

- Content standards and systems

The DSA requires intermediaries to be transparent about any restrictions in their terms and conditions "that they impose in relation to the use of their service in respect of information" provided by users, including information on "any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review" (Article 12). Larger platforms must also set out clearly their policy regarding the suspension of users for frequently providing manifestly illegal content (itself a requirement of the DSA), as well as the suspension of processing of complaints from complainants that frequently submit manifestly unfounded complaints (another DSA requirement) (Article 20(4)).

The Indian Rules require intermediaries to publish prominently on their websites or mobile applications their rules, regulations, privacy policy and user agreement, and any changes to those rules (sections 3(1)(a) and (f)). They must also inform users at least annually that they may terminate the user's account for non-compliance with those documents (sections 3(1)(c)). As noted above, details about the complaints system must also be made prominently available (section 3(2)).

The Brazilian Marco Civil requires "clear and full information" to be included in service agreements, including details concerning the protection of connection and access records and traffic management practices (Article 7(VI)), "clear and complete information" on the collection, use, storage, processing and protection of personal data (Article 7(VIII)), and clear information about any terms of use (Article 7(XI)).

As noted above, the Canadian Technical Proposal requires OCSPs to "publish clear content-moderation guidelines, applicable to the five (5) types of harmful content", which may also be prescribed by the Commissioner (clause 13).

- Recommender systems

Pursuant to the DSA, where very large platforms use recommender systems, they shall set out in their terms and conditions, "in a clear, accessible and easily comprehensible manner, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters that they may have made available" (Article 29).

- Reporting obligations

The DSA imposes a general obligation on all intermediaries other than those which are micro or small to publish, at least once a year, "clear, easily comprehensible and detailed reports on any content moderation they engaged in" during the reporting period. These must at least include information on orders received by States, including to act against illegal content or provide information, on complaints, including responses to them, on *suo moto* content moderation, again including measures taken, and complaints via the internal complaint-handling systems that large platforms are required to establish (Article 13). Additional annual reporting obligations apply to larger platforms, mostly based on their additional general obligations (for example in relation to out-of-court dispute settlements and suspensions of frequent abusers, noted above), but also in relation to automated means of content moderation (Article 23). Very large online platforms need to publish these reports every six months and are again subject to more onerous reporting obligations based on their additional obligations (for example about risk assessments and mitigation measures and audit reports and the implementation reports for them, on which see below) (Article 33).

Where hosting services remove or disable access to content, the decisions and the statements of reasons they are require to provide to users (noted above) must be published in a publicly-accessible online database which shall be managed by the Commission (Article 15(4)).

The original (2017) NetzDG already placed extensive semi-annual reporting obligations on platforms which receive more than 100 complaints per year in its Article 1(2), which need to be published not only online on their own websites but also in the official gazette. In addition to details about complaints and their processing, the reports must cover issues such as general efforts by the platform to eliminate criminal activity on their network, information about submitting complaints and the resources available to the units responsible for processing complaints. These obligations were expanded significantly in the amendments to the law.

The Indian Rules require significant social media intermediaries to publish monthly compliance reports describing all complaints received, actions taken in response to them, the number of takedowns pursuant to their own monitoring and "any other relevant information as may be specified" (section 4(d)). Similarly, news publishers and on-demand providers are

required to publish monthly reports on complaints received and action taken on them (section 18(3)). These actors and self-regulatory bodies are also required to "make true and full disclosure", updated monthly, of all grievances received, how they were disposed of, any action taken on them, the reply sent to the complainant, any orders or directions they received under the Rules and the action they took in response to those orders or directions (sections 19(1) and (2)).

The Brazilian Fake News Law imposes extensive reporting obligations on applications. Pursuant to Article 6, they must publish electronically, in Portuguese and updated weekly, information about: the number of posts against which measures were taken, their location, the grounds for the measures and the methodology used to decide on the application of those measures; the number of artificial disseminators, artificial disseminator networks and sponsored content which were removed or suspended, along with the reasons, location and how they were detected; the number of measures against content which were reversed; and comparison of removal of content and accounts in Brazil with other countries and over time. For users who were removed, disaggregated data on gender, age and location must be provided.

Pursuant to Article 7, the following information must be published quarterly, or weekly during elections: number of accounts in Brazil and number of active users; the number of inauthentic accounts removed, classified by type of inauthentic behaviour; the number of artificial disseminators of content and instances in which non-sponsored content was removed or limited; the number of complaints about illegal, inauthentic behaviour, along with the source of and reason for the complaint; the time taken between receipt of complaints and a response being provided; interactions with disinformation including number of views, shares and complaints, the latter broken down by type of person (natural, legal or official); and the institutional measures put in place to combat disinformation, including a comparison between Brazil and other countries.

According to Article 17, private messaging services must provide the same information as is required by Articles 6 and 7, to the extent that this is technically possible.

The Canadian Technical Proposal requires OCSPs to provide reports, "on a scheduled basis", to the Commissioner containing Canada-specific data about a wide variety of content issues. These include "the volume and type of harmful content" on their services, "the volume and type of content that was accessible to persons in Canada in violation of their community guidelines", "the volume and type of content moderated", the "resources and personnel allocated to their content moderation activities", "their content moderation procedures, practices, rules, systems and activities, including automated decisions and community guidelines", "how they monetize harmful content" and, when relevant, information about their reporting on unlawful content to law enforcement (clause 14).

- Ads

Under the DSA, large platforms which display advertisements need to make it clear to users in real time and in a clear and unambiguous fashion what information is advertising, on

whose behalf the advertising is displayed and the parameters used to determine how the advertisement got displayed to the user (Article 24). Very large platforms need to compile and make publicly available a repository containing detailed information about all advertisements – including the content of the advertisement, on whose behalf it was displayed, the period for which it was displayed, whether it was targeted at certain groups and what parameters were used for that purpose, and the total number of recipients reached – which should be maintained until one year after the advertisement in question was last shown (Article 30).

The Indian Rules require significant social media intermediaries which carry advertising to make that information clearly identifiable to users as advertising (section 4(3)).

The Brazilian Fake News Law imposes extensive transparency obligations in relation to what it describes as sponsored content (content shared in exchange for a payment or other value). Article 7(VIII) requires applications to publish, at least quarterly, reports on who paid for sponsored content, who the target audience was and how much was spent. Chapter III focuses exclusively on transparency in relation to sponsored content. Article 19 requires applications to provide all users with easily accessible information about all of the sponsored content they were exposed to over the previous six months. All sponsored content should be labelled in a way that: identifies it as paid content, indicates who paid for it, provides the contact details of the sponsor, outlines the criteria used to determine the target audience, and provides data on all of the content the sponsor has paid for in the last 12 months (Article 20). Social media platforms should also disseminate, on an unrestricted access website, data on all active and inactive sponsored content relating to "social, electoral and political issues" (Article 23).

5.2. Impact Assessments

The DSA imposes a number of due diligence and mitigation obligations on very large platforms. These platforms should undertake an annual assessment of any significant systemic risks relating to dissemination of illegal content, negative impacts on the human rights to privacy, freedom of expression, freedom from discrimination and the rights of the child, and intentional manipulation of their service, "including by means of inauthentic use or automated exploitation of the service", which have a foreseeable negative impact on issues like health, civic discourse and elections. This assessment should also take into account how their content moderation, recommender and advertising systems impact any of these systemic risks (Article 26). Where the relevant Digital Services Coordinator or the Commission so requests, these platforms must provide vetted researchers (who meet certain conditions) with access to data for purposes of conducting research into these systemic risks (Article 31(2)). They then need to put in place "reasonable, proportionate and effective mitigation measures" for these risks, potentially including adapting their content moderation or recommender systems, their decision-making processes or their terms and conditions, limiting advertising, reinforcing internal processes and so on (Article 27). They also need to conduct audits to assess compliance with various DSA obligations, in particular those set out

in Chapter III. Strict conditions are set for who may perform these audits and for the audit reports that must be produced. The platforms are then required to "take due account of any operational recommendations addressed to them with a view to take the necessary measures to implement them" and to produce an audit implementation report on the measures taken (Article 28).

The United Kingdom Online Safety Bill still relies quite heavily on broad risk assessments in relation to protection of children (see section 10 of the January 2023 version of the Bill) and used the same approach previously when it addressed harm to adults through legal content (now removed, as noted above).

5.3. Media and Information Literacy

The laws providing for regulation of online platforms and other online actors do not tend to focus heavily on educational or awareness-raising issues, which are likely reflected more in funding and action plans of relevant bodies. However, where these laws create oversight bodies, they often do include awareness-raising among their functions. As an example, the Australian Online Safety Act includes, among the functions of the Commissioner, the following:

> [T]o support, encourage, conduct, accredit and evaluate educational, promotional and community awareness programs that are relevant to online safety for Australians (section 27(1)(f)).

The Act also includes a number of provisions on awareness as possible examples of the types of issues that might be addressed in the industry codes and standards that it provides for. A general example of this is placing an obligation on relevant actors to promote "awareness of the safety issues associated with social media services".[77]

It is also likely that some of the more sectoral pieces of legislation include more specific awareness-raising obligations. As an example, Article 45 of the Spanish Organic Law for the comprehensive protection of children and adolescents against violence requires public administrations:

> [To] develop education, awareness and dissemination campaigns aimed at children and adolescents, families, educators and other professionals who regularly work with minors on the safe and responsible use of the Internet and information and communication technologies, as well as the risks associated with the inappropriate use of them which may lead to sexual violence incidents against children and adolescents such as *cyberbullying*, *grooming*, gender-based cyberviolence or *sexting*, as well as the risks associated with access to and consumption of pornography among the underage population.[78]

They must also make available a helpline to provide advice to the same actors (Article 45(2)).

[77] Section 138(3)(l). See also sections 138(3)(m), (n), (r), (s) and (t).
[78] Article 45(1). Note that this is an informal translation of the text.

In Brazil, the State's constitutional duty to provide education shall be understood to include awareness-raising on the "the safe, conscious and responsible use" of the Internet (Article 25 of the Fake News Law).

6. Human Rights Assessment and Recommendations

This section of the Paper provides a broad assessment of the human rights implications of the various approaches outlined in the preceding sections. It does not provide a detailed human rights assessment of the different approaches taken. Instead, reflecting the very diverse nature of the systems reviewed, as well as the dynamic and challenging nature of trying to address online harms through legal regulation, it adopts more of a pros and cons assessment approach, highlighting the relative strengths and weaknesses of different approaches from a human rights, and particularly freedom of expression, point of view.

The first section of this Paper outlined how the volume of harmful online content might change the calculus of the necessity part of the analysis of the test for restrictions on freedom of expression, suggesting that this might justify systemic measures against such content even where it was not legitimate to ban individual instances of it. Despite this, by and large, the systems reviewed here that have been put in place or are being proposed by democracies do not reference the volume issue or appear to use it to justify the measures they impose. Instead, they mostly focus either on measures to address illegal content, which falls outside of that hypothesis, or the taking down of legal content as opposed to more systemic measures.

One exception to this, albeit not specifically justified on that basis, is the risk assessments or due diligence reports that some systems require regulated entities to conduct, as well, albeit to a lesser extent, the compliance audits that, again, some systems require. In better case scenarios, risk assessments cover a number of areas of risk, including rights such as freedom of expression, equality, fair elections and privacy, other social values such as health, and also malign forms of behaviour such as inauthentic accounts or other means of manipulating the platforms. Where these are coupled with requirements to address the risks exposed by the assessments, which they generally are, this can provide a flexible and potentially powerful way to respond to precisely the sorts of volume-based, systemic harms that platforms have engendered. As such, this approach is strongly recommended.

For this to work, the conduct of the assessments, the response to them in terms of mitigation efforts and the oversight system all need to be robust. The oversight system, in particular, needs to ensure that assessments are independent and vigorous. It is likely that this will require robust oversight to ensure that independent researchers conduct the assessments and that they have very broad access to the systems and data of the platforms, even where this is private in nature or includes sensitive competitive commercial information. More research on how this could be achieved is needed.

For this to work it is also essential that genuine efforts are made to respond properly to the risks which the assessments bring to light. The latter needs to be the case even where what is required undermines the engagement-driven business model of the platform, which is likely

to be often, is otherwise costly, again often likely, or can be expected to meet resistance from the platforms. This will require a strong and well-informed stance on the part of the oversight body.

Another important human rights issue is the enhanced enforcement, through many of these systems, of the rules on illegal content, i.e. through engaging regulated entities in taking down or blocking access to it. A first issue here is that where the system relies on content rules which are themselves not human rights compliant, expanding their enforcement reinforces that human rights abuse. Many of the rules in question are indeed not human rights compliant, even in established democracies. And the breadth of some of these rules – like the DSA which effectively applies to all content which attracts liability – exacerbates this problem. A possible solution here is to focus only on more problematical types of content such as serious criminal content which tends to be spread widely digitally, rather than covering all illegal content.

A second issue here is how liability of the regulated entities is engaged. There are, as was noted in the first section, problems with a notice and takedown approach, given that regulated entities cannot be expected to determine, reliably, what content is legal or illegal and, to avoid the risk of being held liable, they are very likely to be overinclusive in taking action against the content. A possible solution here is to require regulated entities to take action when put on notice about possibly illegal content but not to penalise them for making the wrong decision, should that ultimately be determined, whichever way their decision was wrong (i.e. whether they left up illegal content or took down legal content). A notice and notice approach, whereby regulated entities were first required to try to locate the person responsible for disseminating the contested content and to let them decide whether or not to defend it, could be another way to limit over-extensive takedown of content.

As far as legal content goes, some systems envisage a role for political actors in determining the scope of this. This can never be legitimate as it offends the rule that restrictions on content need to be set out in law. While some delegation of authority is permitted under this requirement, that should not extend to discretion to determine what is and what is not prohibited.

In terms of measures to address harmful content, the systems reviewed here tend to rely very heavily on takedowns or blocking, although some do envisage a wider set of measures. Given the very heavy-handed nature of takedowns and blocking, and the requirement under international law that sanctions for breaches of restrictions on freedom of expression be proportionate, far more attention and emphasis should be given, especially in systems that address legal but harmful content, to wider possible measures. Similarly, systems which envisage removing repeat offender users should incorporate safeguards to ensure that this is only done in appropriate – i.e. sufficiently serious – cases.

Most of the systems reviewed here are very broad in scope in terms of the actors covered. At a minimum, a sort of triage approach, such as is found in the three-tier set of obligations in the DSA, with additional carve-outs for very small providers, should be employed. It might also be important to consider differentiating between different types of providers, since not

all are centrally run and controlled, and not all are driven primarily by making increased profits through boosting engagement (i.e. clicks). Wikipedia, for example, is a large provider, especially if viewed through the lens of non-registered users (i.e. people who actually use it regularly), and yet has a fundamentally different operating model.

Quite a few of the systems reviewed here do include provisions aimed at ensuring that users whose content is affected have protections, for example in the form of fair complaints systems. A challenge here is finding an appropriate balance between affording reasonable due process protections to users and having a system which is sufficiently lean and efficient to be able to handle a large volume of complaints. It may be that more experimentation with these systems, or more research, is needed to find that balance.

Some of the systems reviewed include among their requirements the enforcement by providers of their own terms and conditions of service. While this is somehow reasonable – if these terms justifiably restrict freedom of expression then they should at least be applied equally to all users – it also raises tricky questions about the freedom of expression obligations of providers and, in particular, the extent to which it is legitimate for them to impose restrictions on freedom of expression which go beyond legal requirements. Obviously it is not appropriate for States, via laws, to require providers to enforce terms which themselves breach human rights. Despite frequent references by a range of authoritative, academic and civil society actors to the human rights responsibilities of providers, very little in the way of clear standards has emerged regarding, specifically, their freedom of expression responsibilities. More research, debate and standard-setting work is needed in this area.

Another issue in relation to terms and conditions of service is what should be included, or should be required to be included, in them. Some of the systems reviewed here are getting close to requiring any rules that may affect user-generated content to be in the terms and conditions and for these to be made accessible (not only published but actually accessible in the sense of understandable) to users. This seems like a clearly warranted requirement which should be imposed far more widely.

Finally, it is now beyond doubt that providers have transparency obligations, but there is still a lot of debate about how far this goes. Beyond the issue address in the previous paragraph, one of the more sensitive issues is transparency around how content, including advertising content, gets promoted to individual users. Transparency rules in this area remain rather weak. For example, the DSA, which goes farther than most systems, only requires very large platforms which use recommender systems to set out in their (publicly available) terms and conditions "the main parameters" of those systems. Obviously at some point recommender systems go to the heart of the commercial confidentiality of the systems used by providers, but it is clear that it would be possible to push far beyond just the main parameters of the system without crossing the confidentiality line. Interestingly, the DSA is far more exigent vis-à-vis platforms when it comes to advertising transparency, including for (just) large platforms.

Otherwise, strong general regular (e.g. annual or more frequent) reporting obligations are found in some, but not all, of the systems reviewed and these represent good, human rights compliant practice.

Produced with the support of: