Note on UNESCO's Guidelines for Regulating Digital Platforms Draft 2.0

March 2023

Centre for Law and Democracy
info@law-democracy.org
+1 902 431-3686
www.law-democracy.org

UNESCO has been preparing a set of Guidelines for regulating digital platforms - titled A multistakeholder approach to safeguarding freedom of expression and access to information – with a first version launched in December 2022 and a second version (Draft 2.0), launched in February 2023, shortly before UNESCO's Internet for Trust conference. This Note[1] provides the Centre for Law and Democracy's feedback on the second draft document (Guidelines). It is divided into two main sections, the first providing general comments on the document and the second providing more specific comments on different paragraphs.

General Comments

This part of the Note provides comments on issues which are general in nature in the sense that they are relevant to two or more parts or paragraphs of the Guidelines.

Defining a regulatory system

The Guidelines rely very heavily on the notion of a "regulatory system" for many of its recommendations and comments. It is clear, from reading the whole document carefully, that a public, statutory regulatory body(ies) with specific and mandatory powers sits at the centre of the idea of a regulatory system (see, for example, para. 37 but also many of the following paras.). However, this is never really stated definitively or clearly, while other possible elements of the system, beyond the idea that different types of public regulators may undertake this function in different countries (again see para. 37). Some readers may go

through the whole document without really understanding this basic construct (especially given that the term "regulatory system", while compendious in some respects, could also be confusing in other respects). The definition of "regulatory system" in the final section on "References to terminology" both fails to clarify the essence of what is envisaged here and is not sufficient coming at the end, after the "Resources" section, such that readers may not even advert to it. A paragraph clarifying at least that a public, statutory regulatory body will play a key role in the regulatory system should be added to the section "Approach to regulation" to make clear the broad parameters of what is envisaged.

Para. 21 states that the approach of the Guidelines is one of co-regulation, which is then defined (while an almost identical definition appears in the section on "References to terminology"). Unfortunately, the definition is neither clear nor well aligned with the way this term is used in other contexts (such as the EU Audiovisual Media Services Directive). For example, as currently cast, it has both the State and "self-governing bodies" (not sure if this is supposed to be "self-regulatory bodies") creating and then "enforcing" or "administering" rules.

At its heart, co-regulation envisages State action (legal rules) which sets minimum standards for behaviour and then creates a system which is enforceable but either leaves primary application of the rules to a self-regulatory body (subject to mandatory backstopping by a State regulator) or engages those who are subject to the rules in a primary way in the system of application of the rules. We suggest that the reference to "cooperation between State regulation and self-regulation" in the first sentence of the definition of co-regulation in the section on "References to terminology" be changed to "interaction" or something along those lines, since in many instances co-regulation, including of platforms, while following consultation, is actually imposed. Then, both para. 21 and the definition should capture the idea of the State (whether through legislation or potentially a statutory regulator) setting the overall framework of rules and the application bodies, which may be self-regulatory or more co-regulatory (the second option above), setting subordinate rules under those primary rules. In an analogous fashion, enforcement or application is done by the application body, but subject to oversight by the statutory regulator (or courts), if it fails to measure up to what is intended by the rules. This should again be clarified in the Guidelines.

Protecting rights while "dealing with" content

The Guidelines repeatedly refer to the idea of protecting "freedom of expression and access to information while dealing with content that is illegal and content that risks significant harm to democracy and the enjoyment of human rights". While freedom of expression is clearly part of democracy and human rights, this somehow gives the impression that there is a conflict or tension here between freedom of expression and harmful content. That is to some extent true, but there is another aspect of this which is never elaborated upon in the Guidelines, namely that some content online can be seen as a direct attack on freedom of expression in one of two ways. First, when this content hammers public interest speakers, it has the impact of diminishing their voices, to the detriment of the freedom of expression

rights of both these speakers and everyone else. Where this is structural in nature – for example through the maelstrom of attacks on women journalists – the impact is also structural and hence magnified – in that case, that we are structurally deprived of important women's perspectives. Second, where information consumers are so inundated with false information that they genuinely cannot understand what is true and false, their right to receive information is interfered with or restricted. It is important for the Guidelines to at least advert to these ideas, including the fact that freedom of expression protects both speakers and listeners. Even if this were done in a footnote, it would enrich the document.

Content which is illegal under international law

In a few places, the Guidelines refer to content which is "illegal under international human rights law". Perhaps this is just a stylistic error but international human rights law does not directly outlaw content apart from in very limited circumstances, such as under Article 20 of the International Covenant on Civil and Political Rights (ICCPR),[2] which bans hate speech and propaganda for war. Likely what is meant here is content which is allowed to be proscribed under international human rights law. The section on "References to terminology" defines "illegal content" more along these lines but that does not remedy these other references in the text of the Guidelines.

Restricting content: individuals vs. platforms

International law has developed fairly clear rules governing the legitimate scope of restrictions on content disseminated by individuals, for example in areas such as hate speech, defamation and disinformation. However, these standards have been developed in the context of the dissemination of this content by individuals. International law is sensitive to context, especially in relation to the assessment conducted under the necessity part of the test for restrictions on freedom of expression under Article 19(3) of the ICCPR which, among other things, balances the impact of the restriction on freedom of expression against the benefits of the restriction in terms of protecting the legitimate aim. This balancing can change significantly when restrictions are sought to be applied to platforms, where potentially thousands or even millions of individual statements may target an individual or social phenomenon, such as voting or taking a vaccine. The fact that the automated features of many platforms specifically contribute to the multiplication of harmful (but individually legal) statements, in pursuit of profits, also affects the necessity balancing exercise, since it speaks to the relative importance of the freedom of expression interest involved.

So far, there has been little opportunity for international human rights courts or even authoritative human rights actors to weigh in on this and to set out principles for how international standards for restrictions might change as between individuals and platforms. However, this notion is somehow implicit in both the moves by many States to require

[2] UN General Assembly Resolution 2200A (XXI), 16 December 1966, entered into force 23 March 1976, https://treaties.un.org/doc/publication/unts/volume%20999/volume-999-i-14668-english.pdf.

platforms to address otherwise legal content and the constant references in the Guidelines to "content that risks significant harm to democracy and the enjoyment of human rights" (which is juxtaposed with "content that is illegal". While the primary role of the Guidelines is not to elaborate on international human rights standards, it would be useful to set out this idea somehow in the Guidelines, at the very least in a footnote.

Reporting to regulatory system vs. making public

In a number of places, the Guidelines call for platforms to report something to the regulatory system (see, for example, paras. 62, 63 and 67). Presumably most, if not all, of these reports should also be made public, perhaps in some cases subject to limited redactions to protect privacy or commercial confidentiality. It would be good to indicate that in the Guidelines.

Size

Para. 10(a) of the Guidelines calls for bodies in the regulatory system to "identify the platforms by their size, reach, and the services they provide", among other things. Otherwise, however, there is no reference to the issue of regulation actually being calibrated to the size of the platform. More is needed on this crucial point than this rather oblique reference. The size of platforms, however it may be defined by a regulatory system, is an essential element in justifying, both from a human rights point of view (NB the discussion above about the necessity analysis and content restrictions) and as a matter of practicality. To put it differently, it is valid to impose more stringent obligations on larger platforms both because of their outsized impact, including on human rights, and because they can reasonably discharge those obligations. This is reflected in many of the national and supra-national regulatory frameworks that have been put in place or are being considered. To clarify this, the Guidelines should indicate the rationale for increasing obligations with size and note that this should be taken into account in the decision to impose obligations through the regulatory system (the current formulation does not do either of these things).

Language

Language is a specific area where the size issue, mentioned above is relevant. The Guidelines call for the placing of some obligations on platforms in relation to language and content moderation, for example calling for content moderation to take place in the country where the content is published, so as to ensure, among other things, fluency in the language (para. 60), there is only one reference to language in the whole section on Principle 2, transparency, in para. 70(h), which again refers to content moderation (i.e. calling for transparency about the linguistic proficiency of human moderators). In other words, there is nothing at all in the transparency section about the extent to which platforms should make information available in local languages. This is a major lacuna which should be addressed.

There are two references to language under Principle 3, user empowerment. The first is in para. 75, which states that each platform should make available "information about its policies accessible … in all relevant languages". While this is positive, the use of the qualifier "relevant" renders it so general that it is almost meaningless (of course national regulatory systems could set more precise standards here). The second is in para. 82, which states, in part:

> Major platforms should have their full terms of service available in the primary languages of every country where they operate, ensure that they are able to respond to users in their own language and process their complaints equally, and have the capacity to moderate and curate content in the user's language.

This, in contrast, is quite specific. However, lack of clarity is introduced due to the lack of definition of a key element, namely the notion of where a platform operates. Almost by definition, platforms, being online actors, are available in every country of the world. And the larger ones have active users in every, or almost every, country. But it is not clear what sort of threshold would qualify a platform as "operating" in a country (or, for that matter, what constitutes a "primary" language). Without some sort of guidance as to what sort of thresholds would be appropriate here – since it cannot be reasonable to expect a platform to discharge these obligations in languages where they have just a handful of users – even this call fails to provide appropriate guidance.

It is acknowledged that it is not easy to come up with very specific standards here. At a minimum, the Guidelines should explicitly recognise that and indicate that this may need to be addressed through national regulation. But it would be helpful if the Guidelines could come up with more specific direction here.

Terms used to refer to different types of measures

A wide variety of terms are used to describe the different sorts of measures that platforms make take in relation to content. "Moderation" and "curation" are very commonly used but other terms, such as "recommender mechanisms", "take down", "remove", "block" and so on, are also used. Measures in relation to content can be divided into two broad categories, those that operate so as to "push" content at users and those that are employed in response to problematical content (i.e. to demote it in some fashion). It would be useful for the Guidelines to define two terms that it is using as shorthand for these two broad categories of measures. This could be the terms "curation" and "moderation", if that is how they are intended to be understood, but they are not defined as such (or indeed defined at all). Of course where more precise actions are being referenced, other terms should be used, as is already the case.

Money and human rights

A key and difficult issue for regulating platforms is the deep conflicts between their business models and the need to address harmful content. Put differently, it is often precisely the sort

of content that is harmful that the automated systems of platforms boost, since this drives engagement which, in turn, drives profits. This is recognised very obliquely in para. 64 of the Guidelines, which calls on platforms to have in place systems to identify and address cases where their automated systems "result in the amplification of content that risks significant harm to democracy and human rights". It is addressed a bit more directly, but still fairly obliquely, in para. 77, which calls on platforms to "consider how any product or service impacts user behaviour beyond the aim of user acquisition or engagement".

We are not calling for the Guidelines to somehow resolve this very difficult tension. However, it arguably lies at the heart of many of the problems with harmful content online and it seems anomalous for the Guidelines not to say more about it. For example, it should be made clear that the human rights responsibilities of platforms include ceding profits where their business models contribute to human rights abuse (and that part of their corporate social responsibility implies the same where their operations cause harms beyond just harming human rights).

Integrating recommendations across the Guidelines

There are a number of cases where the Guidelines introduce recommendations in relation to specific issues which seem to have more general relevance across the regulatory system. In other words, although these recommendations are certainly relevant in the context in which they are introduced, they would also be relevant to a much wider set of issues that are covered in the Guidelines and sometimes across almost all such issues.

For example, in para. 77 the Guidelines call for thought to be given to how to integrate digital literacy in "all product development teams". Similarly, para. 78 calls for training "all product development teams" on media and information literacy. These are surely ideas that should be recommended far more broadly across the Guidelines (they only appear in these two paragraphs). Paras. 79 and 81 call for partnering with experts in the area of media and information literacy, while para. 95 calls for openness to expert input on risk assessments. Once again, the idea of integrating external expertise is again something that is surely applicable far more broadly across the areas addressed by the Guidelines and yet is only explicitly mentioned for platforms in these provisions. Para. 87 calls for the regulatory system to use its powers to respond where platforms fail to moderate or curate content in accordance with their terms of service or to report fairly and accurately on this to the regulatory system. This again seems to be something that would be far more widely applicable to failures to meet the regulatory standards recommended by the Guidelines. We recommend that these sorts of issues be addressed in a general way early on in the Guidelines. For example the Guidelines could recommend that external expertise be integrated into specific regulatory actions whenever possible.

Specific Comments

This part of the Note provides specific comments directed at just one paragraph, organised by the relevant paragraph.

Paragraph 5:

This claims that the Windhoek+30 Declaration "set three goals to guarantee that shared resource for the whole of humanity: the transparency of digital platforms, citizens empowered through media and information literacy, and media viability". This is a very odd (arbitrary) way of synthesising the main thrust of the Windhoek+30 Declaration. At a minimum, this should be amended to suggest that these were among the goals set by that Declaration, as opposed to THE goals it set.

Paragraph 10(a):

It is recognised that it is very difficult to define the scope of the Guidelines in terms of "platforms" in a clear and precise manner. However, to define this essentially as actors that "allow users to disseminate content to the wider public", with a non-exclusive list of what this might cover following that, is simply too broad. This would cover anyone who operates a website which allows user comments, which would include many NGOs, commercial establishments and public sector bodies. Some additional qualifications are needed.

Paragraph 11(a):

Here, and in one or two other places, the Guidelines refer to their goals as including to support a shared space for stakeholders to "debate and share good practices". While the consultative process of developing the Guidelines does mean that they represent broad, albeit moderated, input (not inappropriately), they do not, of themselves, constitute a debate or offer space for this. If UNESCO intends to provide such a space beyond the February 2023 Internet for Trust Conference, this should be communicated in more concrete terms. Otherwise, these expansive claims for the Guidelines should be removed.

Paragraph 18:

This repeats language from the Rabat Plan of Action regarding a "six-point threshold" for defining criminal hate speech. While it is not unreasonable for the Guidelines to use the language of the Rabat Plan of Action, in fact these are not threshold criteria but, rather, factors to be taken into account in determining whether a particular speech act rises to the level of criminal hate speech. More importantly, it is not clear why, from among all of the many types of speech which are (legitimately determined to be) illegal, hate speech is singled out here for specific treatment. It would be preferable simply to drop this reference.

Paragraph 24:

This refers to "legitimate content" but it would be better to use the phrase "protected content". Legitimate imparts a sense of approval whereas this would encompass, for example, racist or sexist speech which should not be banned by States (but should not be deemed to be "legitimate").

Paragraph 27(f):

This calls for States not to impose a general obligation on platforms to take "proactive measures in relation to illegal content". This is too broad. While platforms cannot realistically monitor content and should never be incentivised to take overbroad measures against content (for example by imposing sanctions on them for leaving up illegal content which they deemed to be legal), ruling out requiring any measures by platforms in this area is too broad. This would, for example, prevent States from requiring platforms to inform law enforcement authorities when they became aware of illegal content (which is legitimate as long as platforms are not penalised for making mistakes in this regard, i.e. for not reporting content which was in fact illegal).

This sub-paragraph includes two entirely different ideas, namely monitoring and proactive measures, on the one hand, and protection against liability in certain circumstances, on the other. These should be separated into two different sub-paragraphs.

Paragraph 27(g):

This protects staff of platforms performing content moderation or curation functions from criminal penalties. While this is considerably narrower than the analogous provision in the earlier public draft of the Guidelines, it is still too broad. It seems to assume that staff are invariably operating in good faith but this is not a legitimate assumption. For example, a racist might get a position working in platform moderation or curation and use that to boost the dissemination of hate speech. Some qualification is needed here, such as that the person did not act with criminal intent.

Paragraph 31:

The first sentence in this paragraph, which is essentially declaratory in nature, suggests that every stakeholder who engages with platforms "has an important role to play in supporting freedom of expression, access to information, and other human rights". That is simply not the case. Think of the racist, malevolent dis-informer or misogynist harasser who is engaging with platforms. This sort of statement is simply too general to be warranted and should be removed.

Section on constitution of the regulatory system:

This section is generally strong and much improved over the previous public version of the Guidelines. However, one important element in protecting both the independence and accountability of regulatory bodies is missing, namely the need for these bodies to have their powers and mandate set out clearly in law, so that any deviations from that could be challenged in court and ultimately reversed.

Paragraph 43:

The opening phrase here refers to "officials or members of the regulatory system". Many of the items that follow are not appropriate for officials (who are likely to be quite numerous,

given the extensive functions of these bodies), such as that they should be appointed on a participatory basis. Instead, the regulator should be able to appoint its own staff (officials). This should be limited to members (or governing individuals).

Paragraph 43(b):

This suggests that members of regulators should be accountable to an independent body such as, among other things, "independent board/boards". There seems to be a bit of confusion here since "members" are normally people who sit on boards of the equivalent thereto. What is needed here is for the entity as a whole, including its govern body, to be accountable to an entirely external body, such as the legislature (already mentioned in this para.).

Paragraph 43(e):

In addition to making conflicts of interest public, there should be clear rules on how they are addressed. These do not need to be spelt out in the Guidelines but the idea of having a system for this beyond just making conflicts of interest public should be referred to.

Paragraph 46(e):

The second part of this, starting with "based on the needs of the public they serve" is not clear.

Paragraph 47:

This calls for a periodic independent review of the regulatory system to be done by a third party reporting to the legislature. The need for this is not clear. Para. 43(b) already calls for the regulatory system to be accountable to an external, independent body, which should review its performance periodically (for example, if the legislature, annually upon submission of an annual report). Normally, taking into account other aspects of the system, including the right of judicial review of decisions, this is enough. It is not clear what sort of third party would be involved here. Of course the accountability function under para. 43(b) might involve the oversight body, such as the legislature, commissioning an independent report or something along those lines where it deemed this to be necessary. But that would be at its discretion and form part of its accountability oversight function (already covered by para. 43(b)).

Paragraph 49:

This calls for decisions limiting content, presumably by the regulatory system, to be subject to being "reviewed by an independent judicial system". It might be useful to distinguish here between judicial and merits review, with the former normally simply assessing whether the regulatory decision was reasonable and the latter assessing whether it was correct, a very different standard. In many cases, judicial review of specialised bodies is limited to the former, respecting their specialised expertise in the area in which they work. This is particularly important in the digital space, where judges often lack even basic knowledge about digital communications.

Paragraph 54:

This refers to applying content moderation fairly "across all regions and languages". It might be useful to add "cultures" here since that is really the most important consideration in this area.

Paragraph 56:

This calls on platforms to "act with due diligence and in accordance with international human rights standards" when they become aware of illegal content. This is very vague and ultimately fails to provide any actual direction as to what is expected. A more specific reference here, at least providing some examples of what a national regulatory system might require to be done, should be added.

Paragraph 57:

There are two problems with this paragraph, which calls on platforms to make illegal content unavailable only where it is illegal and to identify such content consistently with international standards. First, even if a jurisdiction has not specifically outlawed content which is harmful and able to be banned consistently with international human rights standards, there would seem to be no reason not to ban that content in that jurisdiction (taking into account that many smaller jurisdictions are not able to keep up with modern trends on harmful information and that some jurisdictions are not oriented towards protecting certain vulnerable groups, such as sexual minorities). Second, this seems not to take into account the difficult but common situation that is presented where national rules are inconsistent with international human rights standards. Ultimately, it is unclear if this paragraph is asking platforms to refuse to take down content which is illegal in a jurisdiction but this is not supported by international human rights standards.

Paragraph 60:

The part of this paragraph calling for support programmes for content moderators, seems to go beyond the proper brief of the Guidelines. While support measures are no doubt warranted for content moderators, similar arguments could be made for all sorts of human resource support (for example, to address the high stress of programming jobs). It is not clear how such support measures, in contrast to the first part of this para. which calls for adequate training and sufficient staffing of content moderation units, would contribute directly to addressing harmful content, the focus on the Guidelines, as opposed to being part of a wider positive human resources environment.

Paragraph 62:

This refers to the need to address content moderation bias "across different content types, languages, and contexts". It may be useful to add in here the idea of vulnerable groups, which is perhaps the most pronounced form of content moderation bias.

Paragraph 65:

This calls for users to be able "to control the algorithmic curation and recommender mechanisms used to suggest content to them". The reference here to "control" is unclear. It also calls for diversity content options on "trending topics" to "be made clearly available to users". This is not very clear. It might be better to call on platforms to take proactive measures to ensure that users are aware of such options.

Paragraph 67:

This sets out various options for what transparency "can" mean for users. To make this a bit more forceful, this word could be replaced by "should".

Paragraph 70(d):

This calls for transparency in relation to content that is being "removed or blocked". However, this should extend to any measures that demote content. If, as recommended above, a general term is defined to cover this sort of measure, that should be used here.

Paragraph 70(e):

This calls for users to be notified "periodically" about changes in content moderation or curation policies. It would be better to call for users to be notified "promptly".

Paragraph 73:

This calls for researchers to be given access to non-personal and anonymised data. This is too weak. Statistical agencies around the world have in place procedures to give vetted, independent researchers access to non-anonymised private data, under analogous conditions to their own staff. This is entirely reasonable; there is no reason to think that properly vetted researchers would be any less reliable as custodians of this information than staff. Similar arrangements are necessary for platforms, in relation to both personal and commercially sensitive data, so as to ensure some sort of independent review of their operations.

Paragraph 75:

This calls for platforms to "demonstrate how users can report potential abuses of the policies". It should go beyond this to call on platforms to put in place user-friendly and easily accessible systems for such reporting.

Paragraph 83:

This calls for harmful content not to be amplified by automated systems due to lack of linguistic ability. This should also cover a failure to take steps to address harmful content, which is likely a bigger problem in practice in this area.

Paragraph 84:

This calls generally for the rights of persons with disabilities to "be taken into account". This is too generic. One option here is to call on platforms to ensure that their operations are at least WCAG compliant and updated to remain consistent with the latest relevant standards in this regard.

Paragraph 86:

This covers situations where platforms allow use of their services by children. However, in practice it is both difficult and very rare for platforms actually to put in place reliable age verification systems. This reality needs to be taken into account in this para. It also calls for "terms of service and community standards" to be "co-created with a diverse group of children". This hardly seems realistic given that terms of service and community standards on adult platforms are not even co-created with adults. Instead, it might be realistic to call on platforms to use age-appropriate forms of consultation to ensure that children's views are taken properly into account when setting terms of service, community standards and indeed other policies.

Paragraph 89:

This starts out by referring to systems to allow users to "raise their concerns" but then modulates into a system of "appeal". There are important differences between these two ideas. If the goal is to provide for an individual appeal system, this should be stated clearly from the outset.

Paragraph 91:

This calls on platforms to "explain processes for appeal" when measures have been taken against their content. This is too weak. Platforms should be required to put in place accessible, user-friendly appeal procedures and to take effective steps to ensure that users are aware of them, especially, but not only, when specific measures have been taken against those users' content. In terms of measures, the para. refers to a long list of possible measures. Here again it would be useful to have a defined term that refers broadly to all such responsive measures (such as "moderation"). It also calls for platforms to "have processes in place that permit users to appeal" content moderation decisions. This is fine but consideration should be given to calling for something more robust, such as independent appeals mechanisms (likely following an internal appeal).

Paragraph 94:

Sub-paras. (b) and (c) here refer, respectively, to the need for periodic assessments to protect minority users and electoral processes. This seems to imply that these might be separate from the periodic assessments this para. otherwise calls for, whereas it would normally make sense to integrate these as specific elements in general periodic assessments. The Guidelines could present these ideas as specific sub-elements of the main periodic assessment that is being called for. The same also applies to para. 98(a), calling for gender impact assessments (i.e. these should normally also be integrated into the regular annual assessment although there may occasionally be a need to conduct a more specific gender assessment). The idea, in para.

98(b), of integrating independent researchers into gender assessments should also be referred to in relation to the para. 94 assessments and the comments provided in relation to para. 73 should be taken into account here.

Paragraph 102:

The second part of this para. is not very clear.

Paragraph 106:

This refers to the potential of digital platforms having been eroded over "recent decades". Given that Facebook is only 19 years old as of today, this may need to be shortened.