Comments on "Guidelines for regulating digital platforms: A multistakeholder approach to safeguarding freedom of expression and access to information. Draft 2.01"

By Guy Berger, personal capacity.

Cape Town, 6 February, 2023

1. Version 2 improves much on Version 1, in terms of conceptualization and condensing issues into clear principles. Creating international benchmarks by which stakeholders can assess regulatory arrangements at regional or national levels, is a very valuable exercise. As always, however, the elaborating of principles is one thing, but if they do not anticipate the challenges of real situations, the chances of them having impact are limited.

On the ground, much current regulation is state-led, rather than representing a multi-stakeholder movement. This situation correlates with many rules are being drawn up and implemented in ways that threaten to harness platforms as part of the state apparatus – which is not always friendly to freedom of expression. Regulators are weak and also lack independence. (There is also a crying need for prioritization and modularity in focusing regulatory arrangements on particular issues (eg. elections), rather than trying to cover the entire waterfront of issues. Such modularity could be discussed as a helpful approach to "eating the elephant" in the Guidelines).

Commendably, the Guidelines in version 2 do underline the importance of independence in regulatory arrangements. However, taking account of realities, it is very hard to envisage changes here – meaning that this part of the Guidelines will lag, and even undercut the other parts. If key decisions are made by compromised regulators in isolation, then legitimate expression on the platforms could be a casualty.

Thus the Guidelines could do well to encourage a perspective that presents governmental involvement in regulation as one (significant) element, but also as one that optimally works alongside that of other actors in regulation. This would signal the many advantages of having a bigger co-ordinated picture for regulation, even in cases where regulators are problematic in various respects.

Without this, the risk is to encourage replacing platform power with governmental power. In very many cases, these may fail to promote information as a public good but instead end up limiting legitimate expression like criticism of authorities. Proposing that corporate and state powers take account of civil society interests, through institutionalised multistakeholder modalities, is a way forward. It can help protect free expression while systemically countering content that harms human rights and works against information as a public good.

The making of content-related rules by authorities (including regulators), and within this frame by companies themselves, can benefit enormously from meaningful civil society participation. Similarly, the massive work of implementation, monitoring, oversight and review of these rules can also be enriched in the same way. UNESCO has an opportunity in the Guidelines to really highlight the centrality of multistakeholder arrangements in regulatory measures. This would do a lot to aid the insights and arguments of actors working for positive change of the status quo in platform governance that aligns with protection of freedom of expression. In summary, the value of multistakeholder governance as a sine qua non could be better and more consistently presented.

Some points elaborating this argument in relation to specific sections of the Guidelines follow below.

2. Lawful but harmful conceptualization

It is strongly suggested to drop the current distinction between content that is illegal and content that is legal but which the companies should assess as harmful.

Platforms should be free to set their own terms of service in such a way as to prohibit what they see as harmful. But it is a different matter for legal requirements for them to prohibit what is legal content per se. This usurps the rule of law in which a state must takes responsibility for defining what is criminal expression, and it goes against the spirit of the three part test for limiting expression which requires that any restrictions be set in law. A typology may be useful in this regard.

Content	Illegal	Harmful
Category A: FoE problems can arise in cases where local law is not in line with international human rights standards; The Global Network Initiative encourages companies to push back in these cases.	☒	☒
Category B: The criterion of proportionality in human rights standards implies there should be lesser sanctions than case A above (assuming A is a justifiable law).	☑	☒
Category C: Content that is both illegal and harmful calls for platform's priority attention in curation and moderation	☑	☑
Category D. Content that is legal should escape sanction until/unless it reaches a clear (and ideally foreseen) threshold of illegality	☒	☑

The wording in Version 2 of the Guidelines that implies regulatory compulsion for platforms to deal with "lawful but awful" content can be found in several places. For example, clause 59 risks endorsing the privatization of censorship, where platforms are required restrict content in a legal vacuum where such content is not actually outlawed. This has been one of the criticisms of the UK's Online Safety Bill, leading to changes in that draft legislation.

A different way to approach the problem has been proposed in Part 3 of the RIA Working Paper prepared for UNESCO. There, it is pointed out that we can learn from the Rabat Plan of Action against hate speech, which assesses when expression of hatred become prioritized for restriction (and then, applied only in terms of necessity and proportionality). This directs actors to systematically assess content for possible restriction based on context, speaker status, reach of content and imminence of danger among other considerations. In this way, it is possible to conceptualize when lawful

expression becomes amplified to the extent of harming rights and as such thereby crosses the boundary of most jurisdictions' legal regimes (eg. inciting actual violence, discrimination or hostility).

On such a basis, there are two benefits foreseen from the point of view of avoiding breaching the right to free expression:

i. Regulatory arrangements can require from companies that they conduct advance risk assessments which set thresholds for possible danger from an illegality point of view. This would also help to avoid moderation that results in servere prior restrictions on content in the absence of such content beginning to scale and combine to constitute a real and likely illegal harm to human rights.

ii. Indeed platforms could be legally required to act in advance (such as serving warnings to perpetrators), so as to avoid situations where it becomes too late to moderate after a wider narrative grows and fuses to clearly breach legality (eg. actually organizing a violent insurrection). Instead therefore of removing lawful expression, companies can use their technology to assess the factors that underpin a judgement about whether a narrative (combined of many individual posts) may be approaching the point of illegality, and warn participating posters accordingly that they may be getting close to a line. Those who ignore such cautions could have their content sanctioned through labels, deprioritization, limits on likes or sharing, demonetization, etc. When legal limits are overstepped, content should be removed and users engaged in continued violations subjected to deplatforming for a period. Naturally, effective appeal processes are required in terms of administrative justice.

In contrast to the above, by using legal force to compel platforms to police what some may perceive as "lawful but harmful" content, there is a strong risk that the UNESCO Guidelines allow for being weaponised for extra-legal repression of legitimate expression.

Fundamentally, authorities should ensure that different harms are spelled out in law, and articulated in a granular way to the panoply of human rights. (See RIA Working Paper, Part 2). If there is not already a legal protection of such rights, it is this gap that needs correction – not the application of measures that circumvent the law.

It goes without saying that "law" in these framings should be in line with international standards for human rights. Where laws do not meet these standards (eg. are too vague, have disproportional penalties, are selectively applied to critics of government, etc.), it is incumbent on platforms to contest related state requirements for restrictions. Ultimately, platforms would need to consider whether it is too much of a compromise of human rights for them to continue providing services in the applicable jurisdiction if human rights are not respected in local law and its application.

2. Co-regulation

Para 21 says the Guidelines' approach is "co-regulation" between state and "self-governing bodies". This formulation can be misinterpreted.

- "Co-regulation" is typically an agreement between a state regulator and an *industry-based* regulator (which is "self-governing").
- Governmental-linked regulation is when an official regulator sets rules without mediation for targeted "self-governing" companies (better wording would be: "solo-governing", since "self-regulation" usually refers to voluntary industry-based or professional-based bodies). This

form of regulation works directly on all targeted individual entities. It is a modality that is not termed “co-regulation” in most uses of the term.

However, it is unclear in current wording which scenario is being described in the Guidelines.

Critically, the formulation as it is in Para 21 also limits the role of third party institutions to “public scrutiny”, i.e. monitoring, of the direct parties in a regulatory relationship. However, such third party actors can be recognized and welcomed into processes for actual formulating rules for platform companies, as well as into process for deciding on internal rules within platform companies. Further, besides for monitoring and formulating of rules, these external stakeholders can have institutionalised roles to assist with implementation, appeal, oversight and review of rules at both levels – i.e. regarding rules for the sector, and rules within each company.

As currently presented in para 21, the self-styled “co-regulatory” approach does not align with the promise of the title of the Guidelines, which is for a “multi-stakeholder approach”. While indeed, co- and direct regulatory arrangements (in the sense above) can benefit from public scrutiny, that is just a fraction of the potential for good regulation on a multi-stakeholder basis.

Later in the document, under clause 31, it is recognized that multistakeholder value can be included in developing, implementing, evaluating rules and in providing oversight. But this seems like an afterthought, and instead of being integrated into Para 21, the first part stands in contradiction to the second.

In similar mode, under the “references on terminology”, the elaboration of “co-regulation” and self-regulation” can profitably be tweaked. Referencing “codes of conduct” only under “self-regulation” may also be taken to incorrectly imply that such mechanisms are not part of “co-regulation”.

3. Principle 3: Media and Information Literacy.

An initial paragraph could be added here, to locate MIL within a frame of promoting knowledge about the rights to freedom of expression, privacy, equality, etc. This can help ensure that MIL is not seen merely as a defensive strategy for individuals, but one that can also mobilize groups to assert their rights against those breaching them, and also for when platforms fail to protect them sufficiently. This is *partly* (but arguably, insufficiently) noted under clause 81. It merits elevating to the start of this section and strengthening.

4. Additional points:

11.c. This paragraph would do well to drop the qualifier “ensuring alignment where possible”, which weakens the prior content.

15. The phrase “Finally, it describes” should be “Finally, they describe”

19. The reference to “The regulator will expect” may connote a single state-based entity at work. Since there can be a plurality of regulators, including industry-wide self-regulatory instances, and since these arguably should also all involve multiple stakeholders (as per the title of the Guidelines), it would be better to say “Regulatory arrangements will expect…”

27e. It would be appropriate here to first advise that States should refrain from engaging in disinformation themselves, as indeed is the call in several UN resolutions.

28b. Platform transparency could be strengthened by adding “after auditable policies”, the phrase “as well as multistakeholder-agree metrics for evaluating performance”. This is important, since as shown in the RIA Working Paper part 3, the current “prevalence” metrics are less than helpful.

28d. In the list of stakeholders to whom platforms are accountable, it would be important to add advertisers who have a major interest in reducing hate speech and disinformation on platforms, and who also pay the bills (often also of haters and disinformation creators) in current models.

52. The wording here could specify that HR and due diligence considerations are integrated “in a granular way” into all stages. At present, the articulation of human rights jurisdiction and platforms is largely limited to generic lip-service (see RIA Working Paper, part 2).

53. To strengthen the call to be consistent with obligations under the UN Guiding Principles, it would be worth spelling out especially the exercise of human rights risk assessments, which are too rarely done even by those platforms that profess adherence to the Principles.

56. The phrase “child sexual abuse materials or other explicit and severe illegal content” opens the door to subjective judgement about what is “explicit” and what is “severe”. It could be left at “child sexual abuse materials” or have added “live-streaming of acts of terror”.

59. This para could be supplemented by another in the vein of alternatives or complementary steps that governments can undertake, such as regulation to recognize and support decentralized platform options (like the Mastodon protocol on which the EU has significant presence). Other examples are for governments to ensure adequate resourcing for public service media and ending impunity for attacks on journalists (online and offline). Reference can also be made to requiring inter-operability (as in the telecoms industry), which would enhance competition and give users the choice to move to platforms with potentially better curation and moderation regimes without being held hostage in the current walled gardens operated by platforms. Not signalling this issue will be a missed opportunity in agenda-setting, and also convey the misleading impression that regulatory arrangements about content issues in a silo could be sufficient to deal with potential rights-harming content online.

66. The recommended notification of users about “when their content is removed or subject to content moderation and why” could be reconsidered. This provision is in the DSA, but even there it has been critiqued in terms of the massive case load that would be required (eg. covering billions of spam and scam comments that removed each quarter on YouTube.) It may be best to indicate upfront that this provision may well vary with size of enterprize, and with the degree to which there are effective redress procedures for users to appeal against actions. Alternatively, if it is to be realistic, then platforms need to be unbundled into much smaller entities with manageable scale to be able to fulfil this function.

70l. Similar to the above, the breadth of the provision requiring transparency on “practices of advertising and data collection” is very important, but currently worded would be extremely wide-ranging, and hence qualification may be needed.

70j. Important to add to this clause on transparency is information about complaints received from state officials and actions taken.

73. It would be best to add the word “potentially” before “harmful content such as…”. This is because, for example, disinformation is neither automatically nor intrinsically harmful (nor illegal in

terms of international standards, except if used to harm human rights – eg. to reputation or political participation).

76. User reporting – reference is made to gender-based violence and harassment. Here it could be very appropriate given UNESCO as the source of these Guidelines, to consider adding the words: "and to journalists, whistle-blowers and human rights defenders".

86. The issue of identifying children raises many complex regulatory questions of privacy, encryption and bulk monitoring. It is suggested to explicitly signal recognition of the complexities here, while indicating that the balances between issues here will likely vary in different regulatory arrangements and on different content issues.

References on terminology: the phrase "on digital platforms" lists social media networks, search engines, app stores and content-sharing platforms. It is suggested to add in messaging systems (to encompass eg. WhatsApp, Fb messager, Telegram, etc.) which are also used as vectors for hate speech and disinformation, even though they do not algorithmically amplify or rank content and are more difficult to moderate than more open platforms.

=

The opportunity to comment on V.2 is appreciated, and it is hoped that the points above are taken in the spirit they are intended, namely as a constructive contribution to a very important endeavour.