collective wellbeing Carnegie UK

MODEL CODE

A reference model for regulatory or self regulatory approaches to harm reduction on social media

January 2023

Introduction

Carnegie UK's work on social media regulation has underpinned policy formation in the UK. In 2018 we described a new way of reducing harm arising from social media that recognised international human rights. We proposed regulation that focussed not on individual pieces of content but on the systems and processes that companies use to run the business, and to incentivise and distribute content. In the UK that took the form of a statutory duty of care with an independent regulator; in other regimes, such as the DSA, a due diligence process. Since 2018 we have developed our model with victim groups, regulators, governments, parliamentarians, social media companies, civil rights organisations and law enforcement. Our approach has been discussed in the UK, the EU, the G7, the United Nations and in places as far apart as Canada, Germany, Mongolia, Slovakia, Australia and New Zealand.

Carnegie UK has argued that:

- the scale, speed and variety of content on internet platforms make it difficult to regulate content directly;
- that design features, business model and user tools have an indirect impact on content and that platforms have some responsibility for those choices;
- these features can have an affect at each of a number of stages in the content distribution chain (seen as a four-stage model, below);
- a regulatory approach focussed on risk assessment of these features may improve the content environment without excessive reliance on content-removal, and balance the potentially conflicting rights of users through proportionate risk mitigation strategies;
- while individual content domains may be differently affected by particular design features, and different user tools may be helpful depending on context, all content flows through the same distribution chain in a service (so the same questions arise about risk of features, even if the answer might differ across content domains).

On this basis, it seems that a common framework could be developed by reference to the four stage information flow model. The framework would form the basis for a company approach to risk assessment and mitigation. This framework could be deployed across multiple content domains and jurisdictions. In adopting this cross-cutting approach, design-based risk mitigation measures can be seen to have cross-domain – and cross harm - effects. The approach may therefore be more efficient for service providers in tackling specific harms across a range of content domains and -potentially – across jurisdictions.

This approach, which sets down principles rather than detailed rules, is flexible:

- defining a skeleton approach allows a company to develop and apply the framework within its own context (rather than imposing specific technical answers across the

sector) – this may allow for a providers to compete on the basis of their approach to safety, allowing greater choice to users;
- it is future-proofed for similar reasons;
- it allows modular development – so that content domains may be incorporated or not, depending on service provider and requirements of local jurisdiction;
- it can sit as part of formal, legal regulation, be part of self-regulatory initiatives, or sit against international framing and provide a common thread amongst these multiple legal structures (though these different legal structures may be more or less efficient in terms of impact and enforceability).

The model code has many similarities with the UN Guiding Principles on Business and Human Rights (Ruggie) and the OECD Guidance for Multinational Enterprises and would be consistent with those approaches.

The four stage information flow model, which reflects the role of the platforms in creating and influencing the flow of content from their users, comprises the following:
- access to the service and content creation;
- discovery and navigation;
- user response tools; and
- platform response.

Access to the service and content creation includes tools available to users to create content (e.g. filters, nudification apps and mechanisms for labelling content), as well as restrictions (eg limits on frequency of posting) but also includes the user sign-up process and the terms of service for use of the platform. So questions around anonymity, multiple accounts, the acceptability of bot accounts and disposable accounts could all be considered here as well as the adequacy of the terms of service (assessed either against national law or international law standards, as appropriate). The main focus in community standards or terms of service tends to lie on user-facing provisions; advertising content policies should not, however, be forgotten; nor the impact of advertising revenue sharing business models on user content creation.

Discovery and navigation covers all sorts of recommendation tools, and features for organising content such as hashtags and feeds highlighting trending issues, as well as search functions/autocompletes. Advertising delivery systems also fit here, including advertiser sign-up processes (KYC), ad content policy and audience segmentation tools.

User response tools allow the user to curate and adapt the online environment, but this category also includes tools for engaging with content (like buttons for example, or features to facilitate reposting and sharing) as well as the ease of making complaints.

Platform response includes moderation and complaints processes, including any user rights of appeal, crisis protocols and transparency reporting.

At each of the four stages, an intervention could be any one of: an ex ante design choice; the provision of tools or other mechanisms; or content specific responses. For example in terms of discovery a service could choose to optimise for authoritative sources; allow users more control to curate their own feed; or introduce suppression measures related to particular content or speaker.

The reference model

This model code is drafted to sit along international human rights standards as a self regulatory tool. As such it is not phrased, in the main, in mandatory terms. A variant of this code, sitting within national law could be more prescriptive, specifying clearly mandatory requirements. Nonetheless, within this model code there are some principles that are phrased in mandatory language. The approach is based on risk assessment and mitigation so a risk assessment approach itself is mandatory. Further principles identify issues which in our opinion should be

considered in a risk assessment. We also give examples of specific features which are often considered risky.

The nature of this high level, principles based approach means that mitigations cannot be identified here – they will be context specific. An exception to this are steps central to the protection of user rights and which seem relevant no matter the content domain and jurisdiction. Such universal mitigating steps we have expressed to be a mandatory requirement, though the details of implementation may be informed by the risk assessment.

There is not a perfect answer to social media regulation that will fit all countries. A race for a commonality can quickly plummet to only the lowest common denominator which exposes the international disparity in application of the ICCPR leaving many people vulnerable. A modular approach (after Watkins, Ness) using a reference model as set out here could be a basis for international agreement based on common principles. This can provide both help for countries with few resources and the beginnings of a detailed framework for those equipped to regulate.

Professor Lorna Woods OBE, University of Essex
William Perrin OBE, Trustee Carnegie UK
Maeve Walsh, Associate Carnegie UK
January 2023
Contact: maeve.walsh@carnegieuk.org

Company orientation towards reduction of harm

Principle 1: Responsibility, Risk Assessment, Mitigation and Remediation

1. The service provider must have a policy commitment to seek to reduce harm, whether in general or in relation to a set of harms, arising from the operation of their service endorsed by the board and significant subcontractors.

2. The governing board of the service provider should apply the United Nations Guiding Principles on Business and Human Rights, as should significant subcontractors in the supply chain. Large multinational service providers with complex supply chains should comply with the OECD Guidelines for Multinational Enterprises. Particular regard should be paid to risks of harm to media and democratic freedoms.

3. As a foundation for harm reduction activity, service providers must carry out a suitable and sufficient risk assessment of their entire service, including risks arising from the practice of outsourcing responsibilities. In doing so, service providers should engage with relevant experts and organisations representing groups adversely affected by operation of the service. A suitable risk assessment will follow international standards if available or best practice. It shall cover all territories where the service has a non-trivial user base, reflecting their local circumstances accordingly.

4. Where harm reduction is focussed on one or a few specific content domains (eg violence against women and girls), service providers should include a survey of the extent to which the relevant content domain arises and results in harm on its service.

5. Risk assessment must also be carried out in relation to the launch of any new service or new feature. Providers of high harm or very large services should operate a precautionary principle in introducing new features, only gradually increasing their availability while monitoring for harm in dialogue with representatives of victims.

6. The service provider must produce a risk mitigation plan addressing the issues raised in the risk assessment and at least covering the issues covered later in this code (including product testing).

7. The service provider must identify appropriate metrics to assess the appropriateness and success of the mitigation plan (or any part thereof) and use them to assess the effectiveness of the mitigation plan regularly (at least annually) and to update it as appropriate. Metrics should be designed so as to allow comparability across assessment periods.

8. The service provider must remain vigilant at all times to reasonably foreseeable events that could give rise to significant harm such as elections, festivals, sports matches etc or observable yet unforeseen events such as civil unrest, war or severe ethnic tensions and mitigate the harm arising from their services in these contexts. Advances in technology leading to or exacerbating harm will occur and should be mitigated as part of ongoing vigilance. In general the service provider should review the success of the mitigation plan at least annually and revise the plan as appropriate.

9. In reviewing progress, service providers must engage with relevant experts and organisations representing groups adversely affected by the relevant content.

10. Risk assessments and mitigation plans should be recorded, retained for not less than three years and published on the service provider's website in an accessible manner in languages commonly used on the service. The service provider should consider instructing third party audits from independent appropriately qualified auditors.

11. Risks assessments should not assume that users and the way they respond to the service and the content on the service are homogenous. Risk assessments must take into account the characteristics of different groups and the differential impact of the features on them as well as the specific risk of harm to which they are exposed. Specifically, harms arising for children should have a separate risk assessment and mitigation process, informed by General Comment No. 25 of the UN Committee on the Rights of the Child in relation to the digital environment.

12. This first principle is a foundation for principles 2-12. The following principles are applied with reference to assessing and mitigating risk arising from the operation of the services, in all non-trivial geographic markets and including the actions of significant sub-contractors.

Principle 2: Safety by Design

1. Bearing in mind the outcomes of the risk assessment, the service provider must implement appropriate technical and organisational measures to embed safety by design in the development and the running of the service and its features. Safety by design does not mean the elimination of all risks but rather the inculcation of an approach where appropriate choices about understanding, minimising or allocating risk can be made.

2. The service provider must take steps to ensure that the design process takes into account the different characteristics of users, aiming to design inclusively.

3. As part of its risk assessment and mitigation processes, the service provider should carry out or arrange for the carrying out of such testing and examination of its service and business systems (including any advertising or paid content systems) to assess the safety of the service by reference to the harms caused in each relevant content domain by the operation of the service. Testing should include "abusability testing" as well as identifying whether features and tools scale well from a viewpoint of user safety.

4. Testing should include systems and tools for recommendation, content curation and moderation, especially automated tools, but also including user empowerment tools. The service provider should test tools provided by third parties, or ensure that those tools have been adequately tested for safety in the service provider's particular context.

5. At least annually, but informed by the risk assessment process in Principle One and testing, the service provider must review and, where indicated by the review, revise those technical and organisational safety by design measures and/or tools. In reviewing design features, the service provider shall consult with relevantly qualified external experts where appropriate.

Principle 3: Education and Training

1. The service provider must put in place appropriate, updated education and training on harm reduction for all staff and subcontractors involved in the content production and distribution chain. This includes senior executives, designers, developers, engineers, customer support and moderators.

2. Where possible the training should be designed in consultation with independent trusted flaggers and/or representatives of survivors of online harms so as to ensure diversity and inclusion.

3. The service provider should provide staff in section 3.1 above with relevant information, training and support on human rights including the importance of an independent media and democratic voice.

Principle 4: Supply Chain Issues

1. The service provider which outsources any part of its business, including moderation of content, applications, GIFs, images or any other content or tools, including 'safety tech', should ensure that each vendor adheres to safety principles and processes in order to deliver the service provider's Terms of Service or Community Standards.

2. Reliance on outsourced content, features or functions must be included as a fact in the risk assessment and mitigation strategy.

3. People doing outsourced work (such as content moderators) should be protected from reasonably foreseeable harm arising from their task through amongst other things a human rights due diligence process such as that described in the OECD Guidance for Multinational Enterprises.

Access to platform and creation of content

Principle 5: Access to the Service

1. The service provider should ensure, and be able to demonstrate, that its sign-up processes take an appropriate, proportionate approach to the principle of "know your client" (KYC), both in relation to users and advertisers. In particular, insofar as the service provider allows anonymity or pseudonymity, these should be included in the risk assessment (taking into account e.g. the user base, focus of the service) and appropriate mitigating steps for any risks identified implemented. Other aspects of that should be taken into account in the risk assessment include:
    1. the extent to which multiple accounts are permitted
    2. bot accounts
    3. extent to which lack of friction in account creation and availability of multiple accounts allow for 'disposable accounts'.

2. The service provider must make its terms of service (including any privacy policy) and/or community standards visible to would-be users and advertisers before they sign up to the service. The terms of service and/or community standards must be expressed in clear and easy to understand language bearing in mind the comprehension capabilities of groups likely to use the service. This includes providing different language versions of the terms of service and/or community standards appropriate to the territories in which the service is made available. The service provider should have in place expanded guidance explaining their terms of service/privacy policies/community standards (and how these are developed, enforced and reviewed, plus the role of relevant survivors' groups and civil society in developing them). It should ensure that training and awareness tools are readily available to users on the Terms of Service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms.

3. A service provider must have terms of service and/or community standards in respect of its users that are fit for purpose taken against its values, local laws and international human rights.

4. A service provider must undertake regular systemic reviews of its terms of service and/or community standards to ensure that they remain up-to-date, effective and proportionate, and amend them when appropriate (taking into account the findings under Principle 1)

5. To allow a user to make an informed choice when deciding whether to use a service, the Terms of Service should clearly state the risks of harm identified in the risk assessments and the steps taken to mitigate them, including if no steps are taken.

6. A service provider should prompt its users to consider their safety and privacy settings; these features are to be designed appropriately in the light of the risks present on the service. The system must default to the most secure settings.

Principle 6: Creation of Content

1. A service provider should consider the appropriate levels of friction in the content-posting process in the light of its risk assessment – for example prompts about harmful language used in a post; number of posts permitted over a given time period; provision of content wrapper features; more than one click required to repost content.

2. A service provider should consider whether any monetisation or revenue-sharing arrangements with content providers provide incentives for or provide financial support to harmful content, and take appropriate steps to mitigate any such risk.

3. A service provider should include any tools it provides for the creation of content in its risk assessment – this includes but is not limited to bots (including chatbots), bot networks, deepfake or audiovisual manipulation capability, the ability to embed content from other platforms and synthetic features such as GIFs, emojis and hashtags. It should consider implementing oversight on third party tools that it allows to interact with its service.

Discovery and navigation

Principle 7: Discovery

1. The service provider should review their recommender systems, whether in relation to content or to other users to follow, especially their automated systems, so they do not promote harmful content in general or that related to a specific content domain identified as problematic. The service provider should check automated systems for bias (e.g. arising from training data). The service provider should consider the risks of tools/features used for organising content (eg hashtags) and what safeguards should surround their use, for example to prevent terms inciting violence against minoritised groups being used.

2. The service provider should consider the impact of autoplay functions, especially in the context of content curated or recommended by the provider. When a service provider seeks to take control of content input away from the person in this way the provider should consider how this feature might affect a person's right to receive or impart ideas.

3. The service provider should consider whether to provide appropriate information to its users about the accuracy (or otherwise) of information (eg flagging content that has been fact-checked) and should make its policies in this regard available.

4. The service provider must consider how its advertising delivery systems affect content seen by users. In particular, it must consider the circumstances in which targeted advertising may be used and managerial oversight over the characteristics by which audiences are segmented where those segments might be computer or user - generated.

5. The service provider must have terms of service and/or community standards in respect of its advertisers that are fit for purpose taken against its values, local laws and international human rights and should have processes in place to enforce that policy consistently.

6. The service provider must consider the need for explainability or interpretability, accountability and auditability in designing AI/ML systems.

7. For users who are children, the service provider should ensure that Principle 7 is applied to reflect their particular characteristics and vulnerabilities, including their right not to see some information.

Principle 8: Navigation

1. Interface design must adopt a user-centred approach, which takes into account an appropriate (bearing in mind the user base of the service) level of safety; interface design must not manipulate users (no dark patterns).

2. The service provider should consider the impact of autocomplete functions and have systems in place to oversee the process of suggesting autocompletes.

3. If the service relies on personalisation, it should consider how to institute oversight over the segments used for personalisation and have policies in place to identify unacceptable or unethical labels, such as might emerge through automation.

4. The service provider should consider the risks around embedded content from other services and click through to external sites, especially in relation to advertising content.

User response, user tools

Principle 9: User-empowerment Tools

1. The service provider must consider what tools, in addition to content and behaviour reporting tools, are necessary to allow users to improve their control of their online interactions and to improve their safety. These could include:
   a) controls over recommendation tools, so a user could choose for example to reject personalisation;
   b) user-set filters (over words, images, sound, videos or topics);
   c) tools to limit who can contact/follow a user, or to see a user's posts;
   d) tools to allow users to block or mute users, or categories of user (eg blocking anonymous and/or unverified accounts);
   e) Controls for the user over who can and cannot redistribute their content or user name/identity in real time.

2. The service provider should ensure that tools provided following Principle 9(1) are easy to use by all groups of users likely to access the service and take reasonable steps to ensure their prominence such that users are aware they exist.

3. For users who are children, the service provider should ensure that Principle 9 is applied to reflect their particular characteristics and vulnerabilities – in particular 9(1)(c).

Principle 10: Virality

1. The service provider should consider the speed and ease of content transmission. This could include, for example, methods to reduce the velocity of forwarding and therefore the occurrence of harm cross-platform.

2. The service provider should assess the risks posed by any features/tools (eg upvote/down vote; like buttons) provided that encourage users to respond and/or to engage with other user's content.

Principle 11: Reporting and Complaints

1. The service provider must have reporting processes that are fit for purpose, that are clear, visible and easy to use and age-appropriate in design and cover all content and behaviour (whether user-generated, service generated (eg autocompletes) or advertising-based). A service provider should consider whether some forms of complaint (eg harassment; image-based sexual abuse) need specially designed reporting processes.

2. The service provider should allow users and others to complain about unsafe design features that are not 'content'.

3. The service provider must provide the opportunity for non-users who are affected by content or behaviour on the service to report that content and/or behaviour

4. A service provider should record complaints in a sufficiently granular manner to feed into risk assessment review processes. The typology should be developed with survivor representatives.

Platform response

Principle 12: Moderation

1. The service provider's policies must be effectively and consistently enforced in accordance with its detailed policies and further guidance. Such further guidance must be in accordance with national law and international human rights.

2. The service provider must have in place sufficient numbers of moderators, proportionate to the service provider size and growth and to the risk of harm, who are appropriately trained to review harmful and illegal content and who are themselves appropriately supported and safeguarded.

3. Where automated tools are used, the service provider must put in place processes to ensure those tools operate in a non-discriminatory manner and that they are designed in such a way that their decisions are explainable and auditable. Users should be informed of the use of such tools. Machine learning and artificial intelligence tools cannot wholly replace human review and oversight.

4. The service provider must establish clear timeframes or other benchmarks for action against non-compliant content.

5. Action in relation to a complaint must be proportionate to the severity of the harm likely to be caused; content contrary to the criminal law is to be dealt with swiftly. The terms of service should make clearly the nature of any such action and the circumstances in which it would arise, as well as details of any appeals process. Action could include:
   a) Label content as inaccurate/misleading;
   b) Demonetise content;
   c) Suppress content in recommender tools and/or search engines;
   d) Geo-blocking of content;
   e) Suspension of content;
   f) Removal of content;
   g) Non-recommendation of user and/or group as person to follow;
   h) The existence of a strike system;
   i) Geo-blocking of account;

j) Suspension of account;
k) Termination of account.

6. The service provider should have systems of assessment and feedback to the initial reporter and the owner of content that has been flagged and actioned to ensure transparency of decision making. Users should be kept up to date with the progress of their reports and receive clear explanations of decisions taken.

7. The service provider should consider the risk of abuse of complaints processes and put in place appropriate safeguards. It should put in place a right of appeal on all decisions made concerning illegal or harmful content, or content that has been flagged as illegal or harmful content. This system cannot displace user rights to take action before the courts. All users must be given a right to appeal any measures taken against them, whether in full or in part. Users must be able to present information to advocate their position.

8. The service provider should have appeals systems which must take no longer than seven days to assess appeals, except in exceptional circumstances which are unforeseeable and beyond the provider's control (see Principle One for discussion of foreseeable events – such as elections – and unforeseeable ones – such as a war).

9. The social media provider must consider putting in place an appropriate trusted flagger programme that maintains its independence from the service provider and from governments. The service provider should:
   a) ensure trusted flaggers are not used as a sole provider of flagging content;
   b) ensure trusted flaggers are appropriately compensated, while not compromising their independence;
   c) hold regular meetings with members of the trusted flagger programmes to review content decisions and discuss any concerns;
   d) provide support to trusted flaggers who are exposed to harmful content in line with the service provider's support to its own moderation teams.
   e) The service provider should have a crisis response protocol that plans for crises of different types in general and, for foreseeable crisis-types, has methodologies to enable the continued delivery of the service without causing harm in accordance with international best practice. This should include occasions when a government seeks to exert undue influence. All protocols should be drafted in clear and precise language. These protocols should include conflict affected and high risk areas, and processes for identifying and monitoring such areas based on existing classifications (eg OECD States of Fragility) as well as monitoring statements from bodies such as the UN or the International Red Cross. Protocols should be tested before deployment and regularly audited in operation.

Principle 13: Survivor Support and Remediation

1. The service provider must take steps to ensure that users who have been exposed to harmful material are directed to, and are able to access, adequate support in the language victims might use. Support can include –
   a) Signposting and access to websites or helplines dealing with the type of harmful content viewed by the user or witnessed by others who may be affected by the content, even if not the designated target;
   b) Information from, and contact details for, services providing victim support or mental health support after being exposed to harmful materials;
   c) Strategies to deal with being exposed to harmful material.

Principle 14: National Law

1. The service provider must have in place a point of contact for law enforcement authorities for each nation in which the service operates. The contact is responsible for

giving information about criminal content to law enforcement authorities in accordance with national law and international human rights (including but not limited to privacy).

2. This information includes –
   a) Information about the content;
   b) The details of the user, including location;
   c) Details of enforcement action on the content undertaken by the provider; and
   d) Other materials relevant to criminal investigations.
3. Information requested by law enforcement authorities in accordance with the law should be delivered with the time frame specified in relevant law, or, where national law is silent and the time frame is reasonably practicable, in the request.
4. Effective protections should be put in place by the service provider to ensure flagging and court orders are not used for malign purposes to remove content deemed objectionable, by government agencies or law enforcement of any kind, nor powerful individuals which is neither contrary to the law nor to the Terms of Service.