How to achieve a delicate balance – principles for regulating internet intermediaries

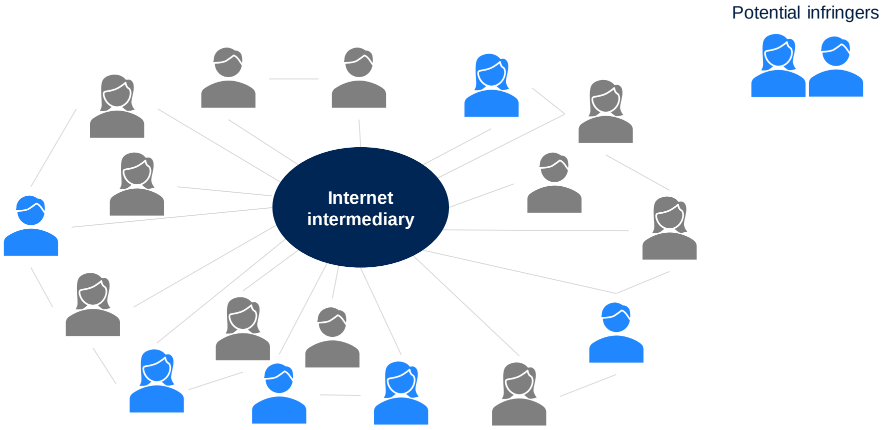
Competition for Future Leaders 2021

1 Introduction

The internet has changed our lives and will continue to do so. It helps us connect with friends and family, study, decide what to buy, discover new hobbies, delivers us the latest news and opinions, gives us quick health and finance information and provides many and varied sources of entertainment. Among many other benefits, it has also given voice to many people and initiatives, and has amplified messages that would otherwise not receive the required time on traditional media. However, it can be a double-edged sword that amplifies harm, disinformation or illegal content and activities. Like never before, global audiences are only a few clicks away. This has prompted regulators across the world to consider how to promote safety and prevent illegal and harmful content online.

Many of our online activities are enabled by internet intermediaries that facilitate the exchange of information between a multitude of parties by transmitting or hosting content. They range from internet service providers, to cloud services, to messaging providers, web hosts, to online platforms. Their unique position and ability to coordinate many different parties at a low cost has unleashed unprecedented growth and solved the coordination problems that dominated pre-internet times. As such, their special position at the centre of activity, means that it could be easier, cheaper and more effective to hold them responsible and request changes through them than through the individuals using their services (see Figure 1).[1]

Figure 1 Connections enabled by an internet intermediary



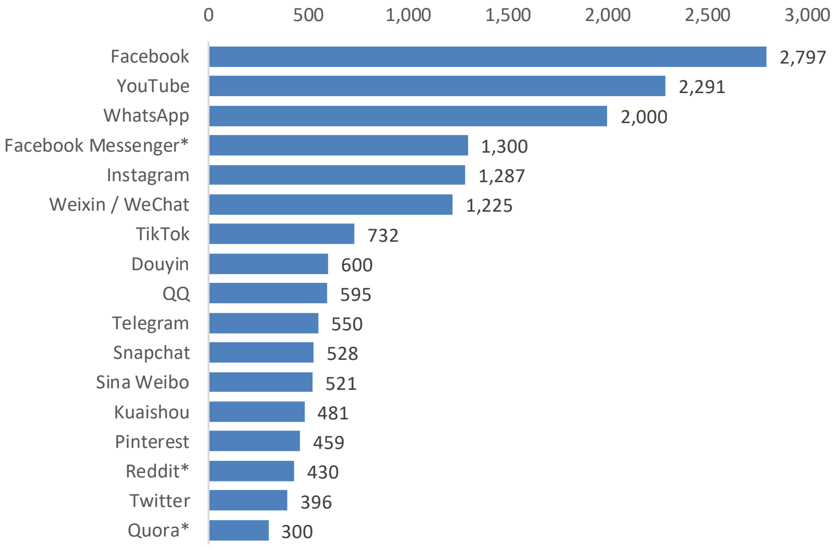
Source: the author.

Regulators across the world are looking into changing the rules that apply to online intermediaries in the context of tax law, competition law or content moderation. This paper focuses only on the issues related to the latter. To understand why this is highly relevant, Figure 2 shows that the most popular social media networks such as Facebook, YouTube or WhatsApp have up to almost three billion users – affecting almost a third of the world's population. This ranking also highlights the diversity of the content that is shared on social media ranging from

[1] Policy Department for Economic, Scientific and Quality of Life Policies (2020), The functioning of the Internal Market for Digital Services: responsibilities and duties of care of providers of Digital Services, requested by the IMCO committee of the European Parliament.

text, to pre-recorded videos, live streams, images, or audio content. All this indicates that the issue of content moderation is broad.

Figure 2 Number of users by social media network (millions)



Source: Statista (2021), Most popular social networks worldwide as of April 2021, ranked by number of active users.
Note: *Platforms have not published updated user figures in the past 12 months.

As policy-makers consider different approaches to assign more responsibility to internet intermediaries, this paper highlights seven principles that can guide them in designing local and international approaches in the evolving communications environment. In short, this paper proposes:

> A regime based on limited intermediary liability with no general monitoring, with additional requirements that are proportionate and target processes not outcomes while safeguarding users, rooted in collaboration and characterised by transparency.

This paper is structured as follows:

- Section 2 defines in:
  - Section 2.1 the spectrum of issues related to content online; and
  - Section 2.2 the trade-offs that arise from intervention in the internet intermediaries’ activity
- Section 3 outlines the seven principles that should be considered by regulators:
  - Section 3.1: provisions for limited intermediary liability;
  - Section 3.2: no general monitoring requirements;
  - Section 3.3: proportionality;
  - Section 3.4: target processes not outcomes;
  - Section 3.5: safeguards for users;
  - Section 3.6: collaboration between multiple parties; and
  - Section 3.7: transparency.
- Section 4 concludes.

2 Issues and trade-offs

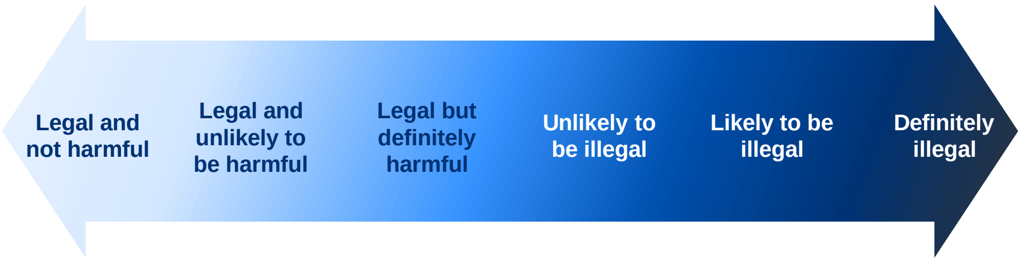
The range of issues that arise online is wide and varied. Policy-makers need to consider first the different categories of issues to be tackled and secondly the trade-offs that exist between the aims they uphold. These are discussed in turn below.

2.1 A spectrum of issues related to content online

Online content can be represented on a spectrum from legal and not harmful to definitely illegal (see Figure 3). The boundaries between the different intermediary categories are not well established. This uncertainty is further compounded by the different national legal systems and cultural factors. At a minimum, the far right side of the spectrum can be established based on what the law considers illegal.

Some of the most common types of illegal content include: child pornography, hate speech, sharing of abhorrent violent material or piracy of copyrighted material. The range of harmful content is much wider and includes, among others, online bullying and abuse, non-consensual sharing of intimate images or videos, advocacy of self-harm, spreading disinformation and misinformation.[2,3]

Figure 3 A spectrum for online content



Source: the author.

This uncertainty in defining the boundaries and different degrees of harm that arise along the spectrum needs to be acknowledged by policy-makers. The same standard of intervention cannot be applied for all issues since this will lead to violations of fundamental rights such as freedom of information or freedom of speech. These trade-offs are discussed below.

2.2 Trade-offs

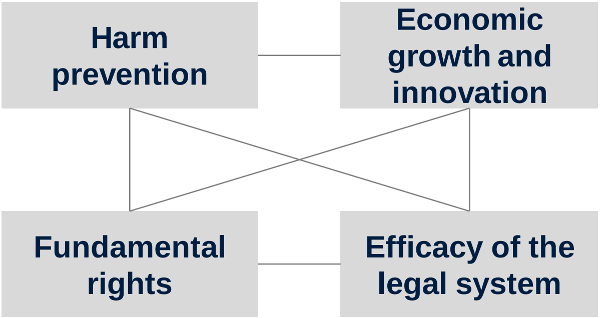
Content moderation online is situated at the intersection of four different aims. Any intervention in the online space will have to account for the trade-offs between the efficacy of the legal systems, harm prevention, fundamental rights and economic incentives for intermediaries (see Figure 4).[4]

[2] UK Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department (2020), Online Harms White Paper: Full Government Response to the consultation Online harms UK.

[3] Australian Government, Online harms and safety.

[4] Daphne Keller (2019), Intermediary Liability: Basics and Emerging Issues.

Figure 4 The trade-offs



Source: the author.

Legal requirements can be set up such that when infringing content is identified it needs to be taken down quickly otherwise the intermediary might suffer financial or legal consequences. However, this can lead to a situation of over-removal that is likely to catch content that is legal and should be protected under the freedom of speech. There is a tension between maintaining online safety and preserving fundamental rights such as freedom of speech, non-discrimination, promotion of democracy and privacy. Furthermore, this can also affect the incentives of online intermediaries which in turn can dampen economic growth and innovation.[5]

The next section will discuss seven principles that can be adopted by policy-makers when designing rules for internet intermediaries that aim to achieve a balance between these trade-offs.

3 Guiding principles

This paper proposes a non-exhaustive list of seven guiding principles for policy-makers when designing approaches for internet intermediaries in the evolving communications environment. These include: provisions for limited intermediary liability, no general monitoring requirements, proportionality, target processes not outcomes, safeguards for users, collaboration between multiple parties and transparency.

3.1 Provisions for limited intermediary liability

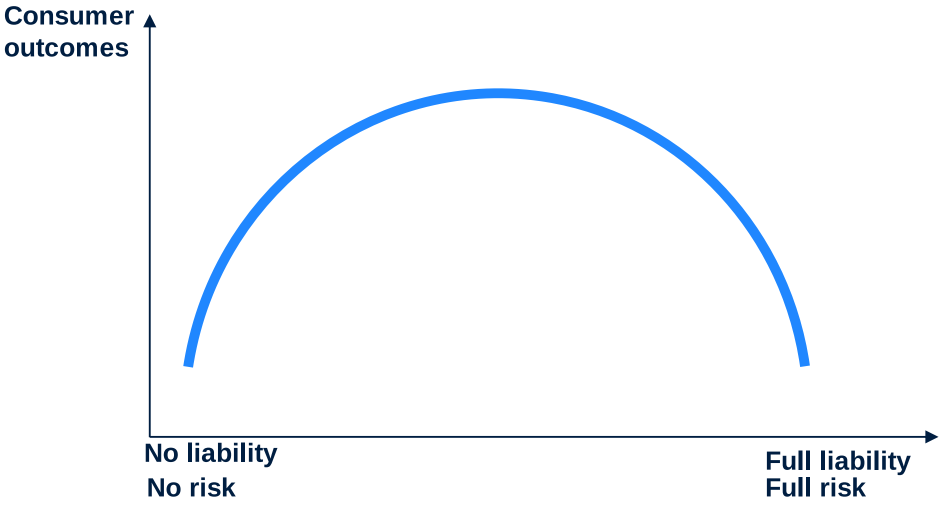
In the introduction we discussed the central role played by online intermediaries and how from a regulatory perspective they are more attractive to regulate than going after individual infringers. While there are good reasons to take this approach, policy-makers should consider how much liability should intermediaries have on a spectrum from none to full liability.

Online intermediaries decide if they want to enter in a market or make investments in innovation if the expected rewards are higher than the costs. The costs are determined also by how much risk they face for being held responsible, the costs associated with compliance with the law and potential damages.[6] A high degree of liability might discourage firms from entering a market or even prompting them to disable certain functionalities on their services – this can have a damaging impact on the services that are based on user generated content. At the other end of the spectrum, no liability would mean low costs for the intermediary but potentially under-enforcement and a high degree of online harm.

A measurement of consumer outcomes that takes into account the benefits from innovation and the negative effects of harms could be maximised at an intermediary level of liability as shown in an illustrative representation in Figure 5.

[5] Pappalardo, K., and Suzor, N. (2018). The liability of Australian online intermediaries. Sydney Law Review, 40(4), pp. 473-474.
[6] The costs associated with liability differ by country. For example, they mainly take the form of damages in the USA, shutdown of the service in Germany or fines and prison sentences in Thailand. See Oxera (2015), The economic impact of safe harbours on Internet intermediary start-ups.

Figure 5 Illustrative consumer outcomes and intermediaries' liability



Source: the author.

The majority of countries around the world have adopted an intermediary liability regime where the intermediaries are protected from liability at least to some degree, however, this is often not sufficiently clear.[7][8] In the EU, the cornerstone legislation that covers the limitations of the intermediaries' liability is the e-Commerce Directive (ECD) which is then complemented by sectoral regulation.[9] This is a type of conditional liability exemption: if an intermediary is hosting, caching or a mere conduit services and complies with the two requirements of (i) not having actual knowledge of the illegal activity and (ii) upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information, then it is protected from intermediary liability.[10] However, the ECD does not establish when and how intermediaries are liable.[11] That is decided at the level of individual Member State or through issue specific legislation at EU level.[12]

On the other hand, the United States has an approach closer to no intermediary liability for 'claims of defamation, invasion of privacy, tortious interference, civil liability for criminal law violations, and general negligence claims based on third-party content' in Section 230(c) of the Communications Decency Act.[13] According to book author Jeff Kosseff, the *Twenty-Six Words that Created the Internet* are:[14]

> "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[15]

This means that in the US, intermediaries have a high degree of immunity for leaving most content online. In addition, there are also provisions for a 'Good Samaritan' protection if the

[7] See for example Alex Comninos (2012), The liability of internet intermediaries in Nigeria, Kenya, South Africa and Uganda: An uncertain terrain or Pappalardo, K., and Suzor, N. (2018), The liability of Australian online intermediaries, Sydney Law Review, 40(4), pp. 473-474.
[8] David Morar and Bruna Martins dos Santos (2020), Online content moderation lessons from outside the US.
[9] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), article 14.
[10] Ibid.
[11] Policy Department for Economic, Scientific and Quality of Life Policies (2020), The functioning of the Internal Market for Digital Services: responsibilities and duties of care of providers of Digital Services, requested by the IMCO committee of the European Parliament.
[12] Swiss Institute of Comparative Law (2015), Comparative Study on blocking, filtering and take-down of illegal internet content.
[13] Adam Holland, Chris Bavitz, Jeff Hermes, Andy Sellars, Ryan Budish, Michael Lambert, and Nick Decoster (2015), NOC Online intermediaries case studies series: intermediary liability in the United States. Gasser/Schulz:, Governance of Online Intermediaries Observations From a Series of National Case Studies, Berkman Center Research Publication.
[14] Jeff Kosseff website, Twenty-Six Words that Created the Internet.
[15] Section 230 of the Communications Decency Act.

intermediary fails to take down content.[16] In terms of copyright, liability is established in the Digital Millennium Copyright Act (DMCA) which also covers formal notices and take down.[17]

In my view, a degree of liability exemption is important, in the sense that it is not only required but also needed, to maximise consumer welfare. The chance of emergence of new categories of harm needs to be considered as an on-going possibility. Furthermore, it must be remembered that innovation may result in both harmful and useful issues simultaneously. As a result, assigning intermediaries all the responsibility has a great potential of sending the wrong signal to the infringers and would possibly limit the potential for innovation and growth because platforms would aim to limit the risk of offering services that can be abused by infringers.

However, only a regime of limited intermediary liability might not be sufficient to solve the issue of preventing harm online while preserving fundamental rights. Where issues persist, the policy-makers can consider imposing additional responsibilities on the internet intermediaries. For example, this is the approach taken by the European Commission through its proposal for a Digital Services Act (DSA). [18] The next six principles can provide guidance on how to set these responsibilities.

3.2 No general monitoring mandate to protect fundamental rights

There are two broad approaches to moderating content online. The first is the ex-post approach where knowledge of infringing content triggers a procedure of take-down. The other is the ex-ante approach where infringing content is prevented from being displayed online from the start. The latter can be operationalised through automated filters and monitoring. However, this creates vast concerns of over-enforcement and violation of fundamental rights such as freedom of speech, freedom of information or privacy.[19]

When a no general monitoring mandate principle is adopted, the intermediaries are free to monitor specific content but cannot be obliged to do so. There are voices that say this is a balance impossible to achieve because automated filtering systems can lead to de facto general monitoring being used by intermediaries.[20] This concern has recently been raised with the new changes to the intermediary liability in India[21] or Thailand where the rules impose in practice a requirement for ex-ante moderation of all content.[22]

One of the solutions for specific filtering is increased transparency that enables visibility over the way in which filtering is done and its outcomes. This is further discussed in section 3.7.

3.3 Proportionality

The next principle is proportionality. The emergence of the internet and the positive network effects that intermediaries facilitate has lead to a large number of companies online. They have different sizes with different business models offering a narrower or wider variety of services. As such, they pose different risks in terms of enabling online harms due to the nature of their activity or the ability to amplify harms online. There are two dimensions on which proportionality can be applied: type of content and risk posed by the online intermediary.[23]

[16] Adam Holland, Chris Bavitz, Jeff Hermes, Andy Sellars, Ryan Budish, Michael Lambert, and Nick Decoster (2015), NOC Online intermediaries case studies series: intermediary liability in the United States. Gasser/Schulz:, Governance of Online Intermediaries Observations From a Series of National Case Studies, Berkman Center Research Publication.

[17] Ibid.

[18] European Commission (2020), Proposal for a 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.

[19] European Digital Rights (2020), Platform Regulation Done Right.

[20] Tambiama Madiega (2020), Reform of the EU liability regime for online intermediaries: Background on the forthcoming digital services act, European Parliamentary Research Service.

[21] Mozilla (2021), India's new intermediary liability and digital media regulations will harm the open internet.

[22] David Morar and Bruna Martins dos Santos (2020), Online content moderation lessons from outside the US.

[23] CERRE (2020), Digital Services Act: Deepening the internal market and clarifying responsibilities for digital services.

Firstly, as discussed above in Section 2, there is a spectrum of issues with online content and not all of them warrant the same level of intervention. With some variations between countries, in practice, the illegal content already attracts stricter responsibilities for intermediaries.

Secondly, the level of risk of a business should be factored in the responsibilities it has assigned. The more is asked from intermediaries in terms of content moderation, ensuring safety and protecting fundamental rights or transparency, the higher the fixed costs each business has. Less burdensome requirements should be placed on the intermediaries with less risk. This will ensure innovation can flourish. Some countries have taken size to be a proxy for the level of risk of an intermediary. This is often a shortcut motivated by the fact that the wider the audience and reach of an intermediary, the more possibilities there are for a larger number of infringements or the amplification of that content to more users. Here is where a concept such as 'public spaces' as opposed to private space, brings to light the importance of the intermediaries services for the citizens and can help separate the large intermediaries that pose a higher risks from those that are large with low risk.[24]

The EU has adopted such a risk-based approach in the proposed DSA. It has created a four-tiered list of responsibilities that increase from intermediary services to hosting services, to online platforms and to very large platforms.[25] Since the cost of compliance might not be proportional to the number of users and the disadvantage this creates for smaller players has determined the European Commission to reduce the number of responsibilities on small firms.

3.4 Target processes not outcomes

The next guiding principle is the targeting of processes in a technology agnostic way as opposed to targeting outcomes.

To begin with, while it might be appealing to policy makers to target an outcome of 'no harm online' this is an impossibly high bar to achieve without radically changing the way we currently communicate over the internet. Such an approach can lead to an extreme outcome for consumers and society where there is no harm online because all services that could enable the dissemination of harms are not available anymore. This would have wide repercussions also on innovation and fundamental rights.

In addition, the types of content that needs to be taken down is evolving and hard rules on specific content that needs to be deleted can quickly become outdated. Putting the pressure on the intermediaries to anticipate this is bound to fail. New ways of circumventing the rules are invented all the time and the regulators might not be able to keep up with the technological developments.

Therefore, adopting as a guiding principle the targeting processes that need to be maintained and used by intermediaries in addition to transparency (discussed in section 3.7) will enable policy-makers to create future-proof regulation. For example, the UK Online harms proposed duty of care framework adopts this as one of the main principles to deal with online harms.[26]

3.5 Safeguards for users

The previous principles focused mainly on how online intermediaries are affected by changes to responsibilities and in liability policy. However, the policy-makers also need to put in place safeguards for users. These can include more visibility (discussed in section 3.7), access to appeals and redress. Without doubt, there will be instances where legitimate content is taken-down incorrectly. When this happens, the users should be able to contest the decision of the

[24] CERRE (2020), Digital Services Act: Deepening the internal market and clarifying responsibilities for digital services.

[25] European Commission, The Digital Services Act: ensuring a safe and accountable online environment.

[26] UK Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department (2020), Online Harms White Paper: Full Government Response to the consultation Online harms UK.

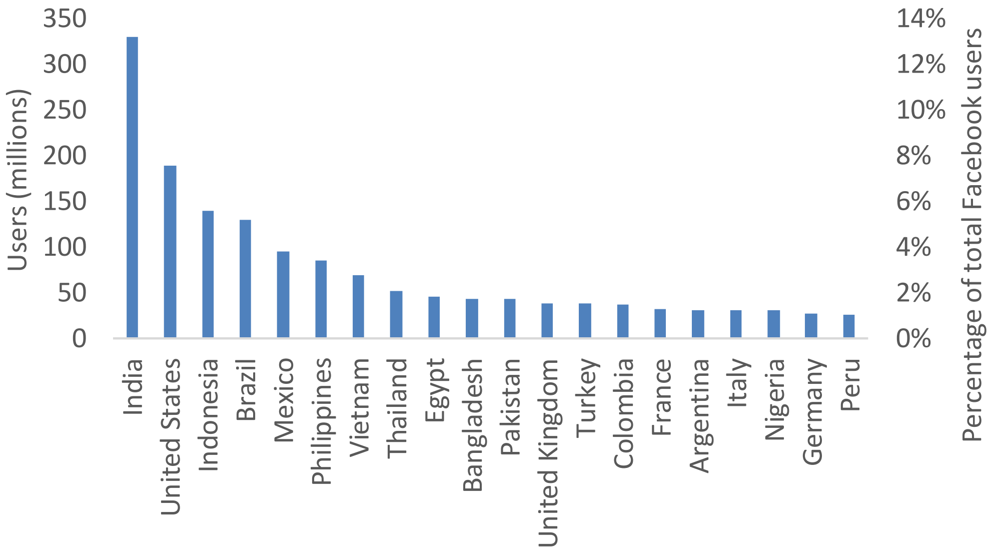
intermediary. This will ensure that there is a balance between the internet intermediary and the users.

While users can appeal through the legal system, this is unlikely to be a viable solution when content moderation takes place at scale. For example, between January 2021 and March 2021, YouTube removed 9.6 million videos of which 93% where detected by automated flagging.[27] As such, it is not surprising that stories of over-enforcement and censorship abound.[28] Faster solutions should be enabled for contesting the removal of legitimate content. Processes such as automatic notifications, easy forms to contact the platform or out of court disputes should be considered where appropriate in addition to the current routes of appeal through courts.

However, it is important to bear in mind that the form these safeguards can take is complex and dynamic. Take for example the Facebook independent Oversight Board that states it aims to "ensure respect for free expression, through independent judgements" and was created to help Facebook answer some of the most difficult questions around freedom of expression online.[29] This board is often called the Facebook Supreme Court given the extensive power it has in deciding what should stay online affecting 2.8 billion users worldwide.[30] This global reach gives this board the ability to influence content moderation to a higher degree than any other individual country however, it is a complex task that needs to account for the differences in culture and national understanding of what is considered legal or illegal or harmful (see Figure 6).[31]

This solution adopted by Facebook is not without criticism. The main worry being the lack of democratic power it holds in making decisions over fundamental freedoms. Take for example the decisions to ban a politician's account or to maintain posts that may enable acts of violence in different countries.[32] This is an area that policy-makers should closely monitor and ensure it is indeed safeguarding users and citizens' interest.

Figure 6 Facebook main users by country



Source: Statista (2021), Countries with the most Facebook users 2021.

[27] Google (2021), YouTube Transparency report.
[28] See for example a regularly updated list by Daphne Keller with Empirical evidence of 'over-removal' by internet companies under intermediary liability laws.
[29] Oversight board website.
[30] Financial Times (2019), Facebook sets out details of 'Supreme Court' for content disputes.
[31] Radiolab Podcast (2021), Facebook's Supreme Court.
[32] The New Yorker (2021), Inside the Making of Facebook's Supreme Court.

3.6 Collaboration between multiple parties

As emphasised throughout this paper, there are multiple challenges that make content moderation complex. The breath of issues that need to be decided and sheer size of content makes it impossible for one entity to enforce. That is where the principle of collaboration can enable more effective rules that also safeguard the interests of the other parties. It requires a departure from the usual regulation approach where one body has all the answers and the market participants need to comply.

Firstly, to achieve collaboration, the policy-makers should aim to align the incentives of the different parties involved so that they can together deliver on a safer online environment. The OECD recognises that self- and co-regulatory initiatives can be a solution:[33]

> Consultation with all interested stakeholders in developing policies can help form the multi-stakeholder partnerships necessary to address complex emerging Internet policy issues.

Collaboration between multi-stakeholder bodies and forums can enable oversight of content moderation at scale. For example, regulators, legislators and courts can provide the framework for content moderation but then be supported by trusteed flaggers, researchers, fact checkers human rights defenders and intermediaries themselves in striking a balance between the trade-offs that arise on content moderation.

One way in which this approach is being implemented is through the use of codes of practice. For example, the EU has a Code of Practice on Disinformation, and Australia introduced the Code of Practice on Disinformation and Misinformation in 2021.[34,35]

3.7 Transparency to safeguard fundamental rights

The final principle, described in this paper is transparency. The ability to scrutinise how content moderation takes place online is vital for allowing researchers, human rights groups, regulators and policy-makers and internet intermediaries to understand what is effective, where are the gaps and how the landscape is evolving. This will inform better solutions. It would also be a solution to monitor the abuse or injustice of content moderation ensuring that it does more good than harm.

Transparency should exist along three dimensions:

- **Explanations on what is displayed online and why**. This will enable users to take more control of their online experience and also become aware of information that is not trustworthy.
- **Audits of the intermediaries' activity**. This can keep tracks of decisions made and aim to discover gaps, emerging trends and level of compliance.
- **Audits of intermediaries' algorithms**. This will open up for scrutiny the black box of algorithms making decisions on content. It will provide transparency on how decisions are made and if its design is flawed in any way. For example, algorithms that are biased against certain groups of users, are incomplete or too interventionist. Access to this might need to be restricted to a small group of researchers or policymakers to protect intellectual property, however, it has the power to enhance fundamental rights protections.

4 Conclusion

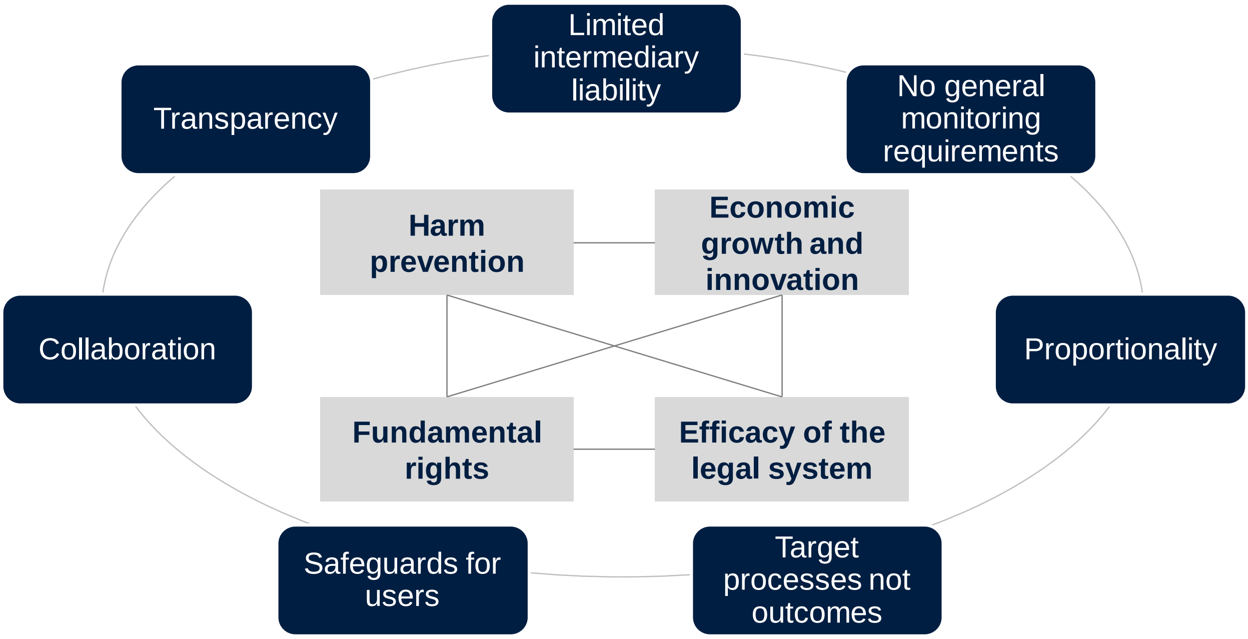
Internet intermediaries play a central and valuable role in today's society. They can enable both great benefits for users and at the same time, they can amplify harms such as the spread of

[33] OECD (2011), The Role of Internet Intermediaries in Advancing Public Policy Objectives: Executive Summary.
[34] European Commission, Code of Practice on Disinformation.
[35] DIGI (2021), Australian Code of practice on disinformation and misinformation.

illegal and/or harmful content such as misinformation or abuse. When policy-makers consider intervening in this space at a local or international level, they need to be mindful of the trade-offs that exist between an efficient legal system, harm prevention, protection of fundamental rights and economic growth and innovation.

In this paper, seven guiding principles have been proposed as guidance to policy-makers to achieve a balanced intervention (see Figure 7). These would set the basis for a regime based on limited intermediary liability with no general monitoring, with additional requirements that are proportionate and target processes not outcomes while safeguarding users, rooted in collaboration and characterised by transparency.

Figure 7 Principles for a balanced approach when regulating online intermediaries


Limited intermediary liability
No general monitoring requirements
Proportionality
Target processes not outcomes
Safeguards for users
Collaboration
Transparency
Harm prevention
Economic growth and innovation
Fundamental rights
Efficacy of the legal system

Source: the author.