A Proposal to Define the Concept of "Online Harmful Content" based on the United Kingdom Experience*

Propuesta para delimitar el concepto de "contenido dañino en línea" a partir de la experiencia del Reino Unido

Jorge Tisné Niemann**

Universidad de los Andes, Santiago, Chile
jtisne@bofillmir.cl | https://orcid.org/0000-0003-2651-2648

◆

Received: June 10, 2024 | Accepted: July 29, 2024 | Published: August 7, 2024

Abstract: The massive use of the Internet has exposed users to a large number of offensive behaviors by third parties, directly violating the rights and freedoms of individuals. In order to promote an online safe environment, different legislative initiatives around the world have begun to tackle online harms, requiring certain moderation duties from platforms. However, the concept of online harm is subjective, especially if we consider that it includes content that is not illegal but is considered harmful or damaging. In view of the above, this article aims to study the different behaviors that can be subsumed under the concept of legal but harmful content in light of the taxonomy of harms offered in the United Kingdom White Paper on Online Harm and then proposes a definition of online harm that serves as a basis for an organic and consistent regulation of the behaviors that must be mitigated and controlled in digital environments.

Keywords: Internet; online harm; legal but harmful behaviors.

Resumen: El uso masificado de internet ha provocado que los usuarios estén expuestos a una gran cantidad de conductas ofensivas por parte de terceros, lo que suele incidir en la vulneración de los derechos y libertades de las personas. Con el objeto de promover un espacio seguro, distintas iniciativas legislativas comparadas han comenzado a regular

* Translated from Spanish into English by Mariano Vitetta.

** LL.B., LL.M. in Legal Research, Ph.D. in Law, all from Universidad de los Andes, Chile. LL.M. in Innovation, Technology, and the Law, Edinburgh, Scotland. Director of the Diploma in Law and Technology, Universidad de los Andes/Eclass. Senior Associate at the IP, Data, and Technology Area, Bofill Mir Abogados.

los daños en línea, exigiendo a las plataformas ciertos deberes de moderación de dicho contenido. Sin embargo, el concepto de daño en línea es subjetivo, especialmente si se considera que incluye aquel contenido que sin ser ilegal, es considerado nocivo o perjudicial. A propósito de lo anterior, este trabajo tiene por objeto estudiar las distintas conductas que pueden subsumirse en el concepto de daño legal pero perjudicial a la luz de la taxonomía de daños ofrecida en el Libro Blanco de Daños en Línea del Reino Unido, para luego proponer una definición de daño en línea que sirva de base para una regulación orgánica y coherente de las conductas que deben ser atenidas y controlados en entornos digitales.

Palabras clave: Internet; daño en línea; conductas legales pero perjudiciales.

1. Introduction

The massive use of the Internet has brought countless benefits for users. In all spheres of life one can see progress in terms of higher access to instant information, tools promoting more effective work, and even the processing of massive data which, together with artificial intelligence, allow to reach knowledge and developments which are expected to redefine the world we live in.

However, together with these benefits, the Internet poses a significant number of challenges which have not yet been addressed appropriately or consistently across multiple jurisdictions. One of those challenges and which, in my view, is one of the most complex to solve is defining what online content should be identified as undesired content and, therefore, prohibited for circulation. This is important especially in societies where the massive use and access to the Internet is a legal interest worthy of protection, even at the constitutional level.[1]

Online harmful content causes serious damaging effects, as it propagates instantly, affects an undetermined number of people, remains visible on the Internet for years and is often generated anonymously. Just like behavior in the offline world, online harm must be addressed appropriately to protect users and ensure that the Internet remains a safe place.

All in all, finding the right balance between undesired content and freedom of expression and information is a particularly difficult task. There is a wide gamut of online harms which are usually indefinite and subjective, and they have to be assessed in connection with equally uncertain elements, such as personal considerations, time of exposition, context,

[1] For example, article 86 of the Proposal for a New Constitution, rejected in the plebiscite of September 4, 2023, enshrined the right to universal access to digital connectivity and communication and information technologies. In that sense, *see* Prince Torres, 2020.

history, and even idiosyncrasy of the individual involved. Moreover, the effort to define a concept of harmful content exceeds the description of merely illegal behaviors, as the most pressing issue, which is the object of this article, has to do precisely with the contents or behaviors which do not violate legal provisions, but which could be considered harmful and undesired for the society and certain individuals in particular anyway.

Defining the concept to be studied in this article is particularly important because the Internet is a highly dynamic environment which is characterized by constant evolution.

Not having a clear concept orienting new regulations trying to control online harmful content creates a scenario in which isolated or fragmented measures may affect technological innovation and freedom of expression, forcing Internet service providers (ISPs) to adopt powerful content-moderation systems, which would likely result in censorship of users. On the contrary, an absolutely unregulated environment, in which all content was allowed, would be detrimental to user participation on the Internet and would undermine any trust in this interaction space.

As in Chile there is no systematic effort to address the issue, it is interesting to be familiar with the initiative by the United Kingdom to control online harm, which started in 2019 with the so-called Online Harms White Paper and which concluded in October 2023 with the publication of the Online Safety Act. While the Online Safety Act ended up imposing major restrictions on regulated online behaviors, the OHWP sparked an intense debate as to the United Kingdom's intent to regulate illegal content and also legal but undesired content (also known as "harmful content"). Part of this basic discussion will be useful to orient the guidelines to be explained in this article.

The OHWP, among other elements, was a pretentious proposal, as its aim was to regulate, from a comprehensive and systematic perspective, a wide list of harmful contents and impose a duty of care for Internet intermediaries or ISPs. Moreover, when dealing with regulations in digital environments, there is usually a lack of a uniform and consistent system grouping such behaviors. This is an issue in Chile, as there is a fragmented regulatory scenario revealing the need to contribute to a definition of the concept of online harm which is useful to guide future public policies on the matter.

Along those lines, some of the behaviors identified as online harmful behaviors are somehow regulated in our country, with scattered regulations. For example, it is possible to identify that sexting, child sexual abuse and exploitation have been included under Law No. 21522 of 2022, amending the Criminal Code. In turn, terrorist content and activity are

regulated under Law No. 18314 of 1984. Incitement to violence is regulated in article 31 of Law No. 19733 on Freedom of Opinion and Information and Practice of Journalism. Moreover, there is now a bill in Congress (Gazette No. 12164-07) with the purpose of punishing revenge pornography. Together with the above, mention must be made of the Advisory Committee Against Disinformation, created under Supreme Decree No. 12, dated May 12, 2023, which came to an end by means of Supreme Decree No. 5 of March 6, 2024, both by the Ministry of Science, Technology, Knowledge, and Innovation.

My claim here is that it is possible to establish the general guidelines of a definition of online harmful content which subsequently allows to conceive an appropriate regulation which is proportional to the control of the content published on digital environments, especially regarding the diligence required from ISPs for their prevention. For methodological purposes, I will use the harms taxonomy under the OHWP as a determined catalog of conducts which may be subsumed in the concept to be defined.

This article is divided into the following sections. First, I will briefly discuss the harms under study and the duty of care provided for under the OHWP. The second section will present the first distinction between illegal contents and legal but harmful contents. The third section will focus on legal but harmful contents, discussing the problems related to their definition. The fourth section will address some elements of hate speech which are proposed as a starting point to define the contours of online legal but harmful content. Based on the above, the fifth section will use those hate speech elements to categorize an ordered and consistent structure of the behaviors identified under the OWHP. Based on the previous analysis, the sixth section will propose a concept of online harmful content.

2. Brief Description of the *Online Harms White Paper* and the United Kingdom's *Online Safety Act*

In April 2019, the United Kingdom's Secretary of State for Digital, Culture, Media & Sport and the Home Secretary jointly presented the OHWP. With this document, the government expressed its intent to be "the safest place in the world to go online, and the best place to start and grow a digital business."[2]

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

In light of this purpose, the OHWP establishes a regulatory framework to address online activity and also undesired online activity. During February 2020, the initial response stage of the OHWP consultation came to an end, and the United Kingdom government answered the main concerns of the interested parties. Afterwards, in December 2020, the government of the United Kingdom published the final and supplementary response on the OWHP. Based on this previous discussion, a bill for an Online Safety Act was submitted to the United Kingdom Parliament, which was passed in October 2023.

The main lines of the OHWP were to provide clarity for companies, a new regulatory framework for online harms, to define the companies subject to the new obligations, powers of the regulatory agency, the compulsory nature of the new regulatory regime, and the use of technology to promote solutions (OHWP, pp. 6-10). For the purposes of this article, I will only address the topic of harms subject to the regulation.

The United Kingdom government was motivated to discuss the OHWP because there is a series of statutory initiatives and other voluntary initiatives by ISPs to address online harms, but they have not been effective enough to stop their propagation on the Internet, especially as to those harms which affect children. The initiative included a wide initial gamut of harms within the scope of the proposal, classified in three different categories: harms with a clear definition, harms with a less clear definition, and underage exposure to legal content.

The list of harms with a clear definition includes: (i) child sexual exploitation and abuse, (ii) terrorist content and activity, (iii) organized immigration crime, (iv) modern slavery, (v) extreme pornography, (vi) revenge pornography, (vii) harassment and cyberstalking, (viii) hate crime, (ix) encouraging or assisting suicide, (x) incitement of violence, (xi) sale of illegal goods/services, such as drugs and weapons (on the open Internet), (xii) content illegally uploaded from prisons, and (xiii) sexting of indecent images by under 18.

The list of harms with a less clear definition includes: (i) cyberbullying and trolling, (ii) extremist content and activity, (iii) coercive behavior, (iv) intimidation, (v) disinformation, (vi) violent content, (vii) advocacy of self-harm, and (viii) promotion of female genital mutilation (FGM).

The list as to underage exposure to legal content includes the following: (i) children accessing pornography, and (ii) children accessing inappropriate material (including under 13s using social media and under 18s using dating apps, and excessive screen time).

As shown, the regulation covered a varied and dissimilar number of types of harms, which rendered the OHWP an initiative which is particularly complex and controversial in the United Kingdom. This was because the aim was to regulate online harms which

could be classified as illegal, but the attempt was also to regulate conducts which did not include a specific definition and, therefore, it would be at the ISP's discretion to prevent those legal but harmful contents.

The OHWP recognized that a relatively small group of companies had participated in the voluntary efforts led by the United Kingdom government to promote online safety. There are different ways in which companies have tried to reduce online harms, such as taking measures once they receive notices of violation of community policies, using moderator teams, or using technology to identify and eliminate any harmful content. But there are lots of concerns about transparency in the implementation of terms and conditions, and there is no consistency, coordination, or general standards among online platforms as to undesired contents and procedures (OHWP, pp. 34-38).

3. The Two-Level Distinction of Online Harms

Addressing online illegal harms has been a constant political concern in Europe and the United Kingdom. An example of this is the Digital Single Market Strategy for Europe (2015), which declared the need to assess the role of intermediaries and online platforms to create a safer Internet (COM/2015/0192, section 3.3.2.). The participation of intermediaries or ISPs has been especially debatable as they "are not publishers, but neither are they neutral conduits; their role in governing online content markets has inevitable ethical connotations" (Bunting, 2018, p. 165).

The OHWP posed the same problem, but adding some controversy as to the concept of online harms. The first distinction that was introduced was that of illegal content and harmful content. In many instances, the OWHP clearly identified the illegal behavior of harmful activities (pp. 5, 11, 15, 42, 44, 66). As obvious as it may seem, illegal harms must be defined in the statute or, at least, the offensive behavior should be covered by a statutory provision to be identified as illegal. According to the European Commission, "illegal content means any information which is not in compliance with Union law or the law of a Member State concerned" (European Commission, 2018, chapter I, 4(b)). On the other hand, harmful activities entail behaviors which are not illegal, but are considered unacceptable.

At the European level, this is not an original distinction in the OHWP, as it had been previously stated, in the response by the government to the Internet Safety Strategy Green Paper (2008), that even if progress had been made in the removal of illegal material

(especially terrorism-related material), further efforts had to be made to reduce the harmful content, both legal and illegal (OHWP, pp. 18-19).

> This classification of two levels of online harms (illegal and legal but harmful) is not peaceful: sometimes the terms "illegal" and "harmful" seem to be used as equivalent terms. For example, the Final Communication from the European Commission on Online Platforms and the Digital Single Market Opportunities and Challenges for Europe (COM/2016/288 final) highlights that there are important areas such as incitement to terrorism, child sexual abuse and hate speech on which all types of online platforms must be encouraged to take more effective voluntary action to curtail exposure to illegal or harmful content. (p. 8.)

It is clear that in the context of that statement when the Commission uses the conjunction "or" in reference to illegal and harmful content, it is assuming that the terms are synonyms, especially because they evoke illegal terms such as incitement to terrorism, child sexual abuse, and hate speech. But the OHWP seems to have departed from this approach, using the conjunction "and" to highlight the distinction, which would admit a clear difference between these terms. This entails a subtle but important distinction, together with the problem of how to appropriately differentiate between both categories.

That being said, it is important to highlight that illegal online harms could not be limited to those listed as clearly defined by the OHWP, as it is expressly mentioned that the list is preliminary, not fixed or exhaustive, based on the predominance of those harms on individuals and the society (p. 30). In addition, the OHWP excluded from its scope some specific online harms (pp. 31-32).[3] However, some authors insisted from the beginning that the OHWP was a "selective, inconsistent, and hierarchical" collection of data, especially because it does not include harms suffered by women and girls (Barker and Jurasz, 2019).

In addition, it is well known that the regulation of online harms is a complex area for public policy, but at least the regulation of illegal harms generates a wider consensus than the idea of regulating legal but unacceptable content.[4] The OHWP also made it clear that the OHWP focused on two illegal harms especially because of their significant impact on individuals and the society (child sexual abuse and exploitation and terrorist activity). Among British authors we have found a similar trend supporting stricter actions and measures for the effective elimination of illegal crimes (related to terrorism, child abuse, and hate speech) (Nash, 2019b, pp. 21-22). However, the OHWP was not exempt from criticism regarding illegal harms, as the vagueness of harms preliminarily listed as clearly defined was criticized. From this point of view, it was stated that OHWP intentionally obscured the

[3] Along these lines, the OHWP excludes all harms to organizations (including the right to competition, intellectual property, and fraudulent activity), harms suffered by individuals in connection with data breaches and online harms suffered by individuals in the dark web.

[4] This was one of the conclusions in a "day-long multi-stakeholder workshop convened to discuss the implications of the 2019 Online Harms White Paper", summarized in Nash, 2019a, p. 3.

distinction to "broaden its scope while making it harder to identify the situations where it is talking about regulating legal content" (Goldman, 2020, p. 359, and note 34). Also, it was stated that it was actually a mixture of harms (Krasodomski-Jones, 2019).[5] As an example and in connection with harms listed with a clear definition, not all types of harassment are illegal and the definition hate crime is not very clear in the United Kingdom (p. 359).

Therefore, even if the OHWP stated its purpose of being the new regulatory framework to improve online safety in the digital economy, it seemed to be a very pretentious proposal, because while it aimed to regulate all harms, it only sketched some.

All in all, given how problematic the first list of illegal activities in the OHWP was, in the complete response to the December 2020 consultation, the United Kingdom government stated that it was necessary to exclude from the new regulation on online content some of the harms which are already covered by other normative bodies, such as those related to (i) intellectual property; (ii) data protection; (iii) fraud; (iv) consumer protection; and (v) cybersecurity or hacking.[6]

Moreover, in the same document, the United Kingdom government defined that legislation would not require the removal of legal but potentially harmful content. However, it was stated that future regulations would try to introduce limitations to other legal but undesired contents with the purpose of ensuring transparency and consistency in the control of online content made by the ISPs (such as contents advocating self-harm, content inciting hatred, online abuse which does not qualify as a crime and contents fostering or promoting eating disorders).[7]

Finally and regarding the discussion on the possibility of limiting freedom of speech by means of prior censorship of online legal but harmful content, the British Parliament decided not to insist on this concept. Given the above, the Online Safety Act focused on the protection of children (including harmful content for children) and the removal of only that which could directly be identified as illegal (including terrorist content). Therefore, after extensive debate, the regulation did not cover legal but harmful content, focusing on primary priority content[8] and priority content[9] which is harmful to children.

[5] Furthermore, Smith (2019) has stated regarding the OHWP that "if the road to hell was paved with good intentions, this was a motorway."

[6] Full response to the consultation on OWHP: https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response#part-3-the-regulator

[7] Idem.

[8] In Chapter 7, Section 61 describes the primary priority content that is harmful to children: (i) pornographic content, (ii) content which encourages, promotes or provides instructions for suicide, (iii) content which encourages, promotes, or provides instructions for an act of deliberate self-injury, and (iv) content which encourages, promotes, or provides instructions for an eating disorder or behaviors associated with an eating disorder.

[9] In Chapter 7, Section 62 describes the priority content that is harmful to children: (i) content which is abusive and which targets any race, religion, sex, sexual orientation, disability, or gender reassignment; (ii) content which incites hatred against people based on the characteristics mentioned in the previous subsection, (iii) content which encourages, promotes or provides instructions for an act of serious violence against a person; (iv) bullying content; (v) content which depicts real or realistic serious violence against a person

4. Examining Online Legal but Harmful Content

After examining the tension resulting from distinguishing illegal content from legal but harmful content proposed in the OHWP, it is now necessary to address the more uncertain and controversial concept of legal but harmful content. We must remember that under the term "harmful content" (or "harms with a less clear definition," the OHWP listed cyberbullying and trolling, extremist content and activity, coercive behavior, intimidation, disinformation, violent content, advocacy of self-harm, and promotion of female genital mutilation (FGM).

The uncertainty regarding this concept has to do with the fact that the OHWP did not define harm or harmful activities, nor did it describe elements which helped to delineate the concept. This means that the term will be defined by the regulatory authority[10] (Smith, 2019b). The initial response to the OHWP did not provide any further clarification on the topic (Smith, 2020). However, it is worth highlighting that the problem associated with a term deficiently defined is not new and has appeared a long time ago as an important issue which is not resolved and must be carefully approached before imposing any regulatory policies (*see* Walker and Akdeniz, 1998, pp. 9-10).

In addition, the OHWP expressly states that the regulatory approach to the legal but harmful content will depend on context and that those harms may possibly include risks related with harms resulting from technology which have still not been identified (p. 42).

Given the wide gamut of harms and the uncertainty as to the concept of harmful content, it has been discussed that, based on the expressions used by the OHWP, harmful content would be anything that is uncivilized, antidemocratic, or harmful to health (Lesh et al., 2019).[11] These concepts, in turn, entail a gamut of behaviors or activities which is too

or which depicts the real or realistic serious injury of a person in graphic detail; (vi) content which depicts real or realistic serious violence or includes a graphic detail of injury against an animal, including fictional creatures; (vii) content which encourages, promotes, or provides instructions for a challenge or stunt highly likely to result in serious injury to the person who does it or to someone else; (viii) content which encourages a person to ingest, inject, inhale, or in any other way self-administer a physically harmful substance or a substance in such a quantity as to be physically harmful.

[10] Regarding the Online Safety Act, the regulator in the United Kingdom is OfCom.

[11] There they ask whether "will we be allowed to criticise politicians with memes? Would humorously edited photos of the Vote Leave's campaign bus come under 'disinformation'? Are questions about the makeup of the UK's immigration intake akin to 'extremist content'?" Similarly, Murray (2019, p. 5) mentions this case which is difficult to solve: "Let's just take one small example of that: imagine that a video-sharing platform hosts a video which shows the aftermath of a drone attack in Syria. Clearly, innocent civilians have been injured. The video mentions Jihad against the United States but mentions no terrorist group. Is this a video showing current events and news, or is this 'terrorist content'? What about if the video showed the aftermath of a bombing in the UK and a commentator mentioned a 'war on terror'? This is but one example

wide to be included in a list. Then, limiting harmful contents introduces an ethical element which is not present in the debate on illegal harms. In addition, it has been criticized that it is a fiction to address legal but harmful contents while protecting freedom of speech, because this freedom will inevitably be limited or restricted during the assessment of the activity's harmfulness (Wragg, 2019, p. 49). Moreover, it has been argued that where the OHWP "fails most seriously" is in trying to regulate in the same manner harmful but legal contents and illegal contents, as harmful but legal contents demand a deeper analysis (*The Guardian*, 2019).

Others have highlighted that the intent to regulate harms beyond the provisions of the criminal code is an encouraging initiative, but it creates the risk that such regulation be used abusively, influenced by means of communication, political opportunism, or by sporadic episodes promoting social discomfort in light of specific situations. Therefore, specifying what is acceptable can be very subjective, and thus the way of restricting freedom of speech online becomes ambiguous (Theil, 2019, p. 44). While harms-based public policies may be welcome, they necessarily require an empirical basis to show how serious the danger to be prevented is. Along these lines, the OHWP did not state any specific arguments in favor of an actual association between the harm experienced or caused with any behaviors qualified as unacceptable. As it is based mostly on presumptions, the initiative was criticized for lacking sufficient grounds to limit freedom of speech (Nash, 2019b, pp. 21-22). Likewise, and due to a lack of solid evidence, it is impossible to predict the consequences of harmful contents, as these consequences will depend on context and the personal perceptions of users (Nash, 2019a, p. 5).

Furthermore, it has been asserted that the way in which the OHWP established the application of the duty of diligence as to legal but harmful behaviors "risks falling foul of the European Convention on Human Rights standards according to which restrictions have to be prescribed by law, and necessary, for a legitimate aim" (Tambini, 2019, p. 33). Therefore, there is a censorship risk when the duty of diligence is decided in codes of conduct prepared by the regulatory authorities, based on harms not defined or with a less clear definition. Any restrictions on freedom of speech must always go through legislative debate and be provided for in the legislation (p. 33).

Madiega (2020) has paid special attention to this point, when insisting on the following:

of a number of very complex value judgments the government seeks to outsource to platforms through the online duty of care."

> a difficult point is that this approach requires distinguishing what is 'illegal content' online from content which is 'harmful' but not illegal, while the concept of 'harmful' is subjective, depends greatly on context and can vary considerably between Member States. Furthermore, fundamental rights defenders argue that introducing rules to address online harmful content into EU law would have grave consequences for freedom of expression, freedom to seek information, and other fundamental rights and therefore seek to strictly limit the scope of the digital services act to illegal content. (p. 11.)

Last, it is appropriate to mention that the OHWP approach was more pretentious than even the national initiatives on online harms, such as the German law on social media (NetzDG), the French law (Proposition de loi contre les contenus haineux sur Internet, 2020) (Cohen, 2019; Hoffmann and Gasparotti, 2020, pp. 25-26), and the Australian amendment to the Criminal Code to punish the exchange of disgusting violent material (2019) (Murray, 2019, p. 3). All these initiatives differ as to substantial elements, but what they have in common is that they mainly address illegal contents (not legal but harmful contents).

An additional precedent of foreign legislation which has tried to address online harmful content is the bill submitted by the government of Canada on February 26, 2024 (Bill C-63). Like the United Kingdom, the purpose of this bill is to create an Online Harms Act. While the denomination of the bill would seem to show that it would also have tried to regulate the harmful but legal content, in reviewing the proposal, one sees that seven types of online harms were prioritized, which also entails amending the Criminal Code of Canada. Along those lines, the bill seeks to prohibit: (i) content that sexually victimizes a child or revictimizes a survivor; (ii) intimate content communicated without consent; (iii) content used to bully a child; (iv) content that induces a child to harm themselves; (v) content that foments hatred; (vi) content that incites violence; and (vii) content that incites violent extremism or terrorism.[12]

It is worth mentioning that the European Union Digital Services Act, which became effective in February 2024, imposes duties of diligence on regulated platforms regarding illegal activities. In that sense, whereas section 12 of that Act is clear when providing

> In order to achieve the objective of ensuring a safe, predictable and trustworthy online environment, for the purpose of this Regulation the concept of 'illegal content' should broadly reflect the existing rules in the offline environment. In particular, the concept of 'illegal content' should be defined broadly to cover information relating

[12] The bill is available at https://www.parl.ca/LegisInfo/en/bill/44-1/c-63

> to illegal content, products, services and activities. In particular, that concept should be understood to refer to information, irrespective of its form, that under the applicable law is either itself illegal, such as illegal hate speech or terrorist content and unlawful discriminatory content, or that the applicable rules render illegal in view of the fact that it relates to illegal activities.[13]

It is interesting to mention that the Digital Services Act imposes on very large online platforms and very large online search engines the duty to adopt proportionate measures to avoid certain risks associated with behaviors which could be harmful to civic discourse and electoral processes, as well as public security, gender-based violence, the protection of public health and minors, and serious negative consequences to the person's physical and mental well-being (article 34).

The above shows that the concern of regulating online harms is increasingly included in foreign legislations. While there is more consensus to control illegal behavior, there is a tendency to incentivize co-regulation with the larger platforms regarding harmful or damaging contents by means of differentiated and proportional control mechanisms.

5. Contribution of Illegal or Undesired Speech as a Starting Point to Assess a Concept of Online Harm

As the notion of legal but harmful content is not precise, any speech or behavior by the users may possibly be considered harmful within the wide gamut of harms under the OHWP (Smith, 2019b). However, with the purpose of defining the concept of online legal but harmful content, it has been claimed that even if "[t]he concept of hate speech is one of the most widely debated yet most elusive in legal studies" (Cavaliere, 2019, p. 6), it seems that the current status of the debate on illegal or undesired discourse could contribute to identifying significant elements for the notion of online harmful content (especially when assessing trolling, intimidation, extremist content, and even disinformation).

Along these lines, it has been claimed that there is a connection between how the regulation of harmful but legal content under the OHWP leads in practice to a regulation of harmful or undesired speech (Haggart and Tusikov, 2019). However, I believe that the term "online harmful content" is wider than "harmful speech" (as the first one limits other

[13] The Digital Services Act exemplifies as illegal behavior the sharing of images depicting child sexual, the unlawful non-consensual sharing of private images, online stalking, the sale of non-compliant or counterfeit products, the sale of products or the provision of services in infringement of consumer protection law, the non-authorised use of copyright protected material, the illegal offer of accommodation services or the illegal sale of live animals (whereas section 12).

forms of harm other than those restricted to the second one). However, the debate on harmful speech seems adequate to distinguish between what is tolerable or not in the digital environment.

Then, given the inherent characteristics of the Internet and of online harmful contents, it is necessary to rethink, identify, and implement new criteria to establish unacceptability thresholds and also to implement adequate technological measures to enforce those thresholds. They must differ from the thresholds traditionally proposed for other areas of the law for what is considered unacceptable.

5.1. Background for the Distinction between Acceptable and Unacceptable Speech in Europe

Freedom of speech has been especially protected in Europe, and the most important provision in that respect is article 10 of the European Convention on Human Rights.[14] One of the first formal documents making a distinction between illegal content and legal but harmful content dates back to 1996, when the Commission of the European Communities published a communication (COM/96/487 final) on illegal and harmful content on the Internet, stating:

> In terms of illegal and harmful content, it is crucial to differentiate between content which is illegal and other harmful content. These different categories of content pose radically different issues of principle, and call for very different legal and technological responses. (p. 10.)

The document emphasizes that the Member States must define what is illegal by law and enforce it by detecting illegal activity and punishing offenders. Harmful contents, instead, are materials which offend the feelings of other persons and must be considered based on cultural and ethical considerations to establish the frontier between what is acceptable or permissible material (COM/96/487 final, p. 11). At the legislative level, Directive (EU) 2018/1808, amending Directive 2010/13/EU (Audiovisual Media Services Directive) sets forth new rules for video-sharing platforms services and social media services as to the protection of children and the general public against harmful contents and incitement to hatred promoted online on video-sharing platform services (articles 6a and 28b) (Montagnani and Trapova, 2019, pp. 6-7). It should also be stated that the Convention on Cybercrime (2001) defines “racist and xenophobic material”[15] and the Additional Protocol

[14] For references on the regulation of freedom of speech in Europe, *see* McGonagle, 2020, pp. 474-480 and Iglezakis, 2017, pp. 367-283.

[15] Article 2 defines the term as “any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, colour, descent or national or ethnic origin, as well as religion if used

to the Convention on Cybercrime (2003) requires Member States to introduce criminal punishments for racist and xenophobic behaviors or for the denial or justification of crimes against humanity (articles 3 through 6).

From a voluntary perspective, the European Union Code of Conduct on countering illegal hate speech online (Code of Conduct) was agreed upon between the Commission and some of the main social media platforms.[16] The Code of Conduct is based on the definition of illegal incitement to hatred included in Framework Decision 2008/913/JAI, of 28 November 2008, focused on crimes related with (i) racism and xenophobia related to publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin; (ii) publicly denying genocide, crimes against humanity, and war crimes; and (iii) denying in a way which could incite to violence or hatred against a specific group of persons based on racism or xenophobia (Article 1(1)).

The Code of Conduct recognizes that illegal incitement to hatred offline is addressed by a robust system of enforcement of criminal law sanctions, but illegal incitement to hatred online requires guidelines so that online intermediaries act promptly against those harms. Therefore, online companies are required to establish their community policies to comply with the provisions in the Code of Conduct.

Even with the legal protection against hate speech, the distinction between appropriate or unacceptable speech seems to be a recurring and gray area of debate. References may be found on this distinction in the Communication of the European Commission on tackling illegal content online (COM/2017/0555), whose purpose is to offer guidelines to online platforms on how to identify and effectively eliminate illegal contents.[17] Scholars have found statements in the document which would suggest "two distinct categories of 'bad' speech: illegal content, defined by national and EU laws, and undesirable content as defined by the platforms themselves" (Cavaliere, 2019, p. 5).

In that sense, the Communication on Tackling Illegal Content Online provides that

> Online platforms should provide a clear, easily understandable and sufficiently detailed explanation of their content policy in their terms of service. These should reflect both the treatment of illegal content, and content which does not respect the platform's terms of service.

In addition,

as a pretext for any of these factors."

[16] The first signatories were Facebook, Microsoft, Twitter, YouTube, and consumer services hosted at Microsoft, such as Xbox gaming services or LinkedIn, to be subsequently joined by Instagram, Google+, Snapchat, Dailymotion, and Jeuxvideo.com.

[17] Such as terrorism, illegal inciting to hatred, abuse of children, or human trafficking.

> [t]he question of whether content is legal or illegal is governed by EU and national laws. At the same time the online platforms' own terms of service can consider specific types of content undesirable or objectionable. (Cavaliere, 2019, p. 16.)[18]

To the statements above I would add the following from the same document:

> There are undoubtedly public interest concerns around content which is not necessarily illegal but potentially harmful, such as fake news or content that is harmful for minors. However, the focus of this Communication is on the detection and removal of illegal content. (COM/2017/0555, p. 6.)

In addition, the document mentions that the European Parliament introduced in 2017 this distinction as it was recommended that platforms reinforced measures to tackle illegal and harmful contents (Cavaliere, p. 2).[19]

The distinction may also be found in the Human Rights Guidelines for ISPs, developed by the Council of Europe in cooperation with the European Association of Internet Service Providers, where ISPs can find guidance on how to address illegal and harmful contents (regarding harmful contents, especially focused on the protection of children) (Council of Europe and EuroISPA, 2008).

All the references reviewed show that there is a political trend to differentiate illegal contents from harmful contents and, in connection with harmful contents, the need to distinguish between what it is undesired or acceptable legal speech. However, the identification of legal undesired speech is not clear; it remains in a gray area which is difficult to solve.

5.2. Elements to Assess the Adequacy of Speech

Hate speech does not cover any kind of speech. Normally, hate speech is connected with extreme intolerance aimed at a specific group. Intolerant speech does not meet the extreme or outrageous standard of hate speech, so it should not be outlawed. In fact, antipathy, disagreement, or intolerance are part of human nature, which in some cases may be even necessary or positive, as in the case of debating about corruption or injustice (Post, 2009, pp. 123 and 125).

[18] I also find precedents for the distinction in European Commission, Recommendation of 1.3.2018 on measures to effectively tackle illegal content online C(2018)1177 final, section 23. http:// data.europa.eu/eli/reco/2018/334/oj

[19] In reference to European Parliament Resolution of June 15 on online platforms (2016/2274/INI).

However, any regulation of speech is particularly complex, as it entails, to some extent, affecting the right to issue an opinion. In that sense, Risso warns that

> there is no doubt that the regulation of hate speech, as well as any limitation of the freedom of thought, entails entering a slippery slope in which it is very difficult to maintain balance. This is especially true when the popular clamor is in favor of a noble cause: putting an end to exclusion, protecting individuals who are being harmed and discriminated against, etc. (Risso Ferrand, 2020, p. 74.)

Post believes that incitement to hatred must be assessed not only in connection with the content of speech, but the way in which it is presented. This refers to the style in which the insult, offense, or degradation is structured. Speech on race, nationality, sexuality, or religion may seem decent and acceptable, but the threshold establishing what is hate speech must be measured by reference to social norms. As to the social norm, the author clarifies the following:

> I shall use the term 'norms' to refer to the group attitudes that we all carry around in us all the time and that form the foundation and possibility of our very 'selves,' and I shall use the term 'community' to refer to the form of social organization that is created and sustained by such norms. (Post, 2009, p. 128.)

Likewise, it has been argued that, as ISPs have great power as to the online interaction of people, they must assume a role and civic responsibility in terms of how they operate their business. According to this point of view,

> Given the international and multicultural contexts in which OSPs operate, the specification of what is socially acceptable and preferable will be effective—i.e. it will be regarded as ethically sound, appropriate, and desirable—only insofar as it will rest on an approach which may reconcile the different ethical views and stakeholders' interests that OSPs face. (Taddeo, 2020, p. 136.)[20]

In addition, four variables have been identified which are subsumed in the different regulations of speech, and each of them is debatable, especially as to the way in which ISPs have enforced them in their community rules (social media). The first has to do with the scope of protection and the group of persons covered by the provisions, which may be based on religion, sexual orientation, disabilities, race, among others. The second variable has to do with the form of speech, in terms of its ability to promote harmful behaviors or

[20] Along the same lines, Helberger et al., 2018, p. 7.

actions (including written, graphic, or video representations). The third variable has to do with the nature of harm, in reference to the relationship between physical and non-physical harms as objects of regulation. While physical harm has traditionally been the object of regulation, non-physical harm requires a reasonableness test. The last variable has to do with the causal link between speech and harm (Cavaliere, 2019, pp. 8-27).

These will be the variables we will use below to delineate and sketch a concept of online harm.

6. Categorization of Legal and Harmful Contents Based on the Elements of Hate Speech

How complex the definition of online harmful contents is has to do with the fact that the concept entails several concepts which are substantially different, especially in the way it has been described under the OHWP. As anticipated, to address this indefinite concept, I will analyze that it is included under the OHWP harms taxonomy in the groups identified as "harms with a less clear definition" and "underage exposure to legal content."[21]

6.1. First Classification: Individual and Social Harms

In addition to the original distinction of online harms (illegal or legal harms), harmful behavior can be classified using the distinction between social and individual harms. It has been claimed that social harms can be clearly defined and their content is objective by nature. However, individual harms are not easily reducible to a defined concept and have to do with concrete facts. Their assessment demands the balancing of subjective elements to establish regulations (Tomlinson, 2019).

This distinction seems to be consistent with the OHWP, when it states that online harms "undermine our democratic values and debate" (p. 5), and under the subheading "Threats to our way of life," the document describes how disinformation and dissemination of inaccurate or false information may be harmful. But in reviewing all the harmful contents listed in the OHWP, there is a clear trend to focus on individual harms rather than on social harms. Therefore, of the contents listed in the OHWP, only disinformation can conform to this classification, leaving the problem open of categorizing for the remaining individual harms.

[21] I want to stress that there are other online harm taxonomies such as cyberharms but not included in the OHWP, as discussed in Agrafiotis, et al., 2018, p. 8. The United Kingdom regulator, OfCom, has also provided their own taxonomy, which can be accessed in OfCom, 2018, p. 12.

6.2. Second Classification: Harms Indistinctly Affecting All Individuals and Those Affecting a Special Group of Persons Based on Sex or Age

Individual harms allow for a new subclassification among unacceptable harms aimed at a special group of people depending on their gender or age and harms affecting all individuals indistinctly.

Of the list of harms in the OHWP, Female Genital Mutilation (FGM) can be classified within unacceptable contents based on gender, but also any other unregulated and unforeseen harmful content, such as sexualized and/or misogynist contents (Barker and Jurasz, 2019), unjustified salary discrimination, or the promotion of eating disorders. It is appropriate to clarify that these contents can also affect men. On the other hand, children accessing pornography and children accessing inappropriate material may be classified as unacceptable contents for their age. This subgroup could also include contents considered inappropriate for older people (as they could be a vulnerable group in the digital environment).[22]

Our decision to use this subclassification and not the variable relative to groups of persons who fit hate speech (based on religion, sexual orientation, disabilities, or any other distinctive element) has to do with the fact that all of them should be included in the term "hate speech," which already exists as illegal harm. However, this opens the door on how to categorize legal but unacceptable speech in the OHWP. We believe that this could be solved using two different strategies: including it as a new concept in the list of "terms with a less clear definition" or subsuming it under an already existing term. I prefer the latter option, subsuming it under "violent content," especially because the OHWP states that "content which is violent with additional contextual understanding" (p. 67)[23] can be considered to be part of the concept of "violent content." We believe this would be the case of legal but unacceptable speech.

6.3. Third Classification: Physical and Psychological Harms

Last, an additional element of speech which may be useful to develop a last subclassification is the nature of harm, because it is based on the physical or psychological impact on the victim. The OWHP only refers to the physical consequences of the harm in

[22] In general terms, the National Consumer Service has identified certain groups of persons as vulnerable consumers in Interpretive Circular on the Notion of Hypervulnerable Consumer dated December 31, 2021. The above has been discussed by López, 2022, pp. 340-415.

[23] Along this line, the OHWP states "Violent content ranges from content which directly depicts or incites acts of violence, through to content which is violent with additional contextual understanding or which is harmful to users through the glamorisation of weapons and gang life."

the cases of terrorism or children (pp. 41-43), but not necessarily for other types of harms (for example, harms which may affect adults, the elderly, or persons with disabilities).

Based on the fact that most harms have been listed as "with a less clear definition," the fragmentation proposed can change depending on the meaning each society assigns to each harm.

This being said, the apology of self-harm and violent contents can be included within physical harms. On the other hand, the rest of uncategorized harms listed in the OHWP (cyberbullying and trolling, extremist content and activity, coercive behavior and intimidation) may be classified within psychological harms.

7. Proposal for a Concept of Online Harm in Light of the Categorization of Harms Proposed

As stated, the problem of conceptualizing the term "online harmful content" derives from the fact that its definition is not in the OHWP or other legal texts and, in turn, it refers to indefinite concepts. Given the above, it is more complex to define a subgroup of harms (legal but harmful content). In other words, if the gender (online harm) lacks a definition, it is more difficult to define the species (legal but harmful content) and, therefore, to categorize each of the subspecies (each harmful content).

The problem with definitions is that they tend to be closed concepts, lacking the necessary flexibility to adapt to the reality they try to describe. Reaching the correct balance between certainty and flexibility is a difficult but necessary effort to offer a clear, useful, and lasting definition, especially for ISPs which need to comply with a duty of diligence regarding such behaviors. In that regard, the interested agents or the individuals possibly subject to duties of diligence identified from a beginning that it was essential that the regulation imposed clearly defined concepts (Center for Democracy & Technology, 2019, p. 2).

Then, a very short definition may exclude new elements (behaviors) which were not foreseen when creating the term. Moreover, a very wide term would not give sufficient certainty regarding its real meaning and what is expected from those who must ensure its enforcement.

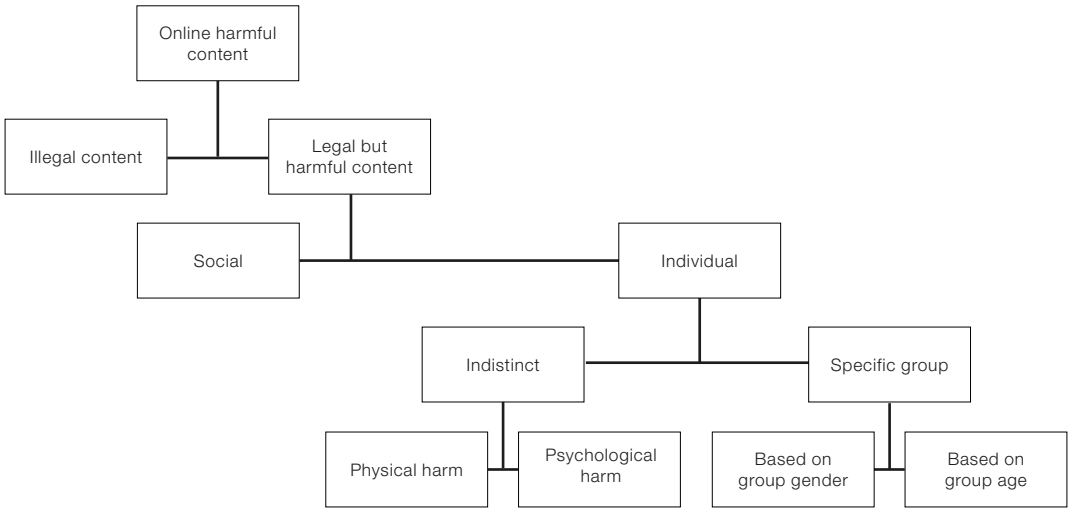
As an example of this approach, "online harm" (which is a wider term than online legal but harmful content) has received a less clear definition as any "behaviour online which may hurt a person physically or emotionally. It could be harmful information that is posted online, or information sent to a person" (OHWP). Likewise, it has been highlighted that legal but harmful content "refers to content which often does not strictly fall under the prohibition of a law, but might nevertheless have harmful effects" (Madiega, 2020, p. 10).

These definitions, other than the distinction whether the content fits or not the law's scope of application, do not provide further details or clarity to the terms.

Therefore, the categorization of the several elements within the term "online legal but harmful contents" is a necessary effort to propose a more objective concept, as it helps understand the object of study, providing at the same time a more detailed approach of each subgroup of undesired behaviors.

As a summary, the following chart (graph 1) summarizes the categorization proposed in this study of online harms:

Graph 1
Proposal to delineate the concept of "online harmful content"



In my opinion, and considering the elements categorized above, the concept of "online harmful content" can be described as follows:

> A behavior occurring in the digital environment, expressed in any form or set of representations, which even if legal may be potentially harmful for a community, a specific group of persons or any individual, based on common and shared belief of what is considered to be unacceptable in a specific democratic society.

Five elements can be highlighted in the notion proposed:

First, it immediately differentiates between legal and harmful contents and illegal contents.

Second, it admits that any representation or set of them (images, text messages, videos, or audio) may express a harmful content, introducing a contextual element for assessment. In turn, this introduces flexibility in the definition.

Third, it distinguishes between social (in reference to the community) and individual harms.

Fourth, it uses generic subgroups (indistinctly affecting individuals or only specific groups), instead of trying to identify specific harms. This also makes the definition adaptable, as any unforeseen harms may be attributed to these or other new groups.

Fifth, the reasons which will explain the threshold between what is acceptable and what is unacceptable cannot be arbitrary, and that is why they must be based on a debate which represents a democratic agreement in a specific society (which introduces the notion of social norms proposed by Post in connection with legal but harmful speech[24]). As the difference is very subtle, the assessment is likely to vary from country to country. However, the demarcation of objects subject to regulation (identified unacceptable contents) will guarantee the protection of freedom of speech and information in every society and, therefore, will allow a more objective regulation regarding the liability demanded from ISPs to control undesired contents.

8. Conclusions

It is undeniable that the Internet has contributed to the development and progress of societies. However, the Internet can also be a source of serious harms for persons. Any online behaviors may affect the rights and freedoms of users, at least to the same degree as if they had taken place in the material world.

After all, determining what must be understood by online harm has not been carefully addressed by local legal theorists and still remains an undetermined concept. The above is not trivial, as a correct understanding of such notion will be useful for the creation of appropriate public policies to prevent and control any content believed to be inadequate or intolerable. Moreover, this definition can be the basis for the creation of consistent and organic systems regarding the wide spectrum of harmful behaviors (currently described

[24] Subjective, contextual, and domestic matters relative to the concept of legal but harmful have been clearly discussed by Madiega, 2020, p. 11.

and new ones arising in the future) and, that way, fragmented answers in the domestic system will be avoided, as has been the case in Chile.

This article has used the debate generated as a result of the United Kingdom OHWP to propose a concept of online harm, as the OHWP was an innovative document which tried to introduce a two-level distinction in terms of online harms, i.e., illegal harms and harms which are legal but harmful.

While the regulation of online harms is an area of constant debate, at least illegal harms enjoy more consensus for their regulation. The above is shown in the online content control laws of Germany, France, Australia, the United Kingdom, the European Union, and the recent bill in Canada which focus mainly on illegal behavior (connected with terrorism, the protection of children, and hate speech).

Conversely, the concept of harm which is legal but harmful introduced in the OHWP is more complex and controversial as it refers to vaguely defined behaviors and which cannot be eradicated in a given set of actions. The risk of regulating these conducts is giving more discretion to ISPs when controlling them, which may entail risks for the freedom of speech of users as they may be subject to previous censorship measures. Therefore, there is a pending debate regarding the tolerance of certain behaviors which, not being illegal per se, are anyway considered undesired. For methodological purposes, the taxonomy offered by the OWHP has been used regarding these harms to identify the behaviors analyzed.

This article has proposed that it is possible to define harm that is legal but harmful regarding harmful speech; there is evidence in other jurisdictions of the need to distinguish from legal but unacceptable speech. Along those lines, we have proposed that the many behaviors which could be limited to legal but harmful harm may be categorized using the distinction between individual and social harms. Then, within individual harms, it is necessary to distinguish between those conducts which indistinctly affect all individuals in a society (admitting a subgroup of physical and psychological harms), and behaviors aimed at specific groups (whether depending on their gender or age).

This classification is important to adopt specific regulations regarding each of the harms categories. Also, their categorization is the previous step to offer a definition of online harm, as behaviors considered in isolation lack any order allowing for their conceptualization.

The above has allowed to fit the multiple behaviors provided for under the OHWP in a clear categorization. Along these lines, the first observation to online harms is that of making a two-level distinction, between illegal harms and harms which are legal but harmful.

Then, illegal contents must contain all those conducts which are clearly defined in the system and which include a specific sanction. Regarding legal but harmful contents,

disinformation would be included within harms affecting the community or the society as a whole. At the individual level, the apology of self-harm and violent content are attributable to possible physical harms on individuals. On the contrary, behaviors of cyberbullying, trolling, extremist content and activity, coercive behavior and intimidation described in the OHWP would be included within the psychological harms affecting individuals. Finally, behaviors affecting a group based on their gender would be the promotion of Female Genital Mutilation (FGM), sexualized contents, misogynist contents, unjustified salary discrimination, or eating disorders. In turn, behaviors affecting a group based on age would be access to pornography by children and access to other kinds of inappropriate materials for children.

Based on the foregoing, and applying the two-phase distinction, as well as the elements which have been contributed to distinguish tolerable from intolerable speech, this article proposes a definition of online harm as the behavior occurring in a digital environment, expressed in any way or set of representations which, even if legal, can be potentially detrimental to a community, a specific group of persons, or any individual, based on shared and common understanding of what is considered unacceptable in a specific democratic society.

About the article

Notes on conflict of interest. The author declares not to have any conflict of interest as to the publication of this article.

Contribution in the article. The author assumed all the roles established in Contributor Roles Taxonomy (CRediT).

References

Agrafiotis, I., Nurce, J., Goldsmith, M., CReese, S. and Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity, 4*(1), tyy006. https://doi.org/10.1093/cybsec/tyy006

Barker, K. and Jurasz, O. (2019). Online harms and Caroline's Law – what's the direction for the law reform? Scripted. https://script-ed.org/blog/online-harms-and-carolines-law/

Bunting, M. (2018). From editorial obligation to procedural accountability: policy approaches to online content in the era of information intermediaries. *Journal of Cyber Policy, 3*(2). https://doi.org/10.1080/23738871.2018.1519030

Cavaliere, P. (2019). Digital platforms and the rise of global regulation of hate speech. *Edinburgh School of Law Research Paper* (2019/29). https://dx.doi.org/10.2139/ssrn.3456141

Center for Democracy & Technology (2019). *Nine Principles for Future EU Policymaking on Intermediary Liability*. CDT. https://cdt.org/wp-content/uploads/2019/08/Nine-Principles-for-Future-EU-Policymaking-on-Intermediary-Liability-Aug-2019.pdf

Cohen, M. (2019). Will the Online Harms White Paper make the UK the safest place in the world to go online? A look at recent approaches the UK, Germany, Australia and New Zealand have taken to regulating online harms. *Computer and Telecommunications Law Review*, 25.

Council of Europe and EuroISPA. (2008). *Human Rights Guidelines for Internet Service Providers*. COE. https://rm.coe.int/16805a39d5

Goldman, E. (2020). The U.K. Online Harms White Paper and the Internet's Cable-ized Future'. *Ohio State Tech. L.J., 16*(2). https://dx.doi.org/10.2139/ssrn.3438530

Haggart, R. and Tusikov N. (2019). What the UK's Online Harms white paper teaches us about internet regulation. *Inforrm*. https://inforrm.org/2019/04/22/what-the-u-k-s-online-harms-white-paper-teaches-us-about-internet-regulation-richard-haggart-and-natasha-tusikov/

Helberger, N., Pierson J. and Poell, T. (2018), Governing online platforms: From contested to cooperative responsibility, *The Information Society*, *34*(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913

Hoffmann, A. and Gasparotti, A. (2020). *Liability for illegal content online Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a Digital Services Act.* CepStudy. https://www.cep.eu/fileadmin/user_upload/hayek-stiftung.de/cepStudy_Liability_for_illegal_content_online.pdf

Iglezakis, I. (2017). The Legal Regulation of Hate Speech on the Internet. In T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (Eds.), *EU Internet Law. Regulation and Enforcement*. Sringer.

Krasodomski-Jones, Alex (2019). Can the government nudge us towards a better internet? *CapX.* https://capx.co/can-the-government-nudge-us-closer-to-a-better-internet/

Lesh, M., Dumitriu, S. and Salter, P. (2019). Safeguarding progress: The risks of internet regulation. Adam Smith Institute. https://coilink.org/20.500.12592/ck82jd

López Díaz, P. (2022). El consumidor hipervulnerable como débil jurídico en el derecho chileno: una taxonomía y alcance de la tutela aplicable. *Latin American Legal Studies, 10*(2), 340-415. https://doi.org/10.15691/0719-9112Vol10n2a7

Madiega, T. (2020). Reform of the EU liability regime for online intermediaries. Background on the forthcoming digital service act. European Parliament Research Service. EPRS. https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA(2020)649404_EN.pdf

McGonagle, T. (2020). Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation. In G. Frosio (Ed.), *Oxford Handbook of Online Intermediary Liability.* https://doi.org/10.1093/oxfordhb/9780198837138.013.24

Montagnani, M. and Trapova A. (2019). New Obligations for Internet Intermediaries in the Digital Single Market — Safe Harbors in Turmoil? *Journal of Internet Law, 22*(7), 3-11. https://dx.doi.org/10.2139/ssrn.3361073

Murray, A. (2019). Rethinking Regulation for the Digital Environment. *LSE Law — Policy Briefing Paper*, (41). https://dx.doi.org/10.2139/ssrn.3462792

Nash, V. (2019a). *Internet Regulation and the Online Harms White Paper Stakeholder Workshop*. https://dx.doi.org/10.2139/ssrn.3412790

Nash, V. (2019b). Revise and resubmit? Reviewing the '2019 Online Harms White Paper' *Journal of Media Law, 11*(1), 18-27. https://doi.org/10.1080/17577632.2019.1666475

OfCom (2018). *Addressing harmful online content: A perspective from broadcasting and on-demand standards regulation.* https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/addressing-harmful-online-content/

Post, R. (2009). Hate Speech. In I. Hare y J. Weinstein (Eds.), *Extreme speech and democracy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548781.003.0008

Prince Torres, Á. C. (2020). El acceso a Internet como derecho fundamental: perspectivas internacionales. *Revista Justicia & Derecho*, *3*(1), 1-19. https://doi.org/10.32457/rjyd.v3i1.45

Risso Ferrand, M. (2020). La libertad de expresión y el combate al discurso del odio. *Estudios Constitucionales*, *18*(1), 51-89. http://dx.doi.org/10.4067/S0718-52002020000100051

Smith, G. (2019a, April 30). Users Behaving Badly: the Online Harms White Paper. *Inforrm*. https://inforrm.org/2019/04/30/users-behaving-badly-the-online-harms-white-paper-graham-smith

Smith, G. (2019b, May 12). The Rule of Law and the Online Harms White Paper. *Cyberleagle*. https://www.cyberleagle.com/2019/05/the-rule-of-law-and-online-harms- white.html

Smith, G. (2020, February 20). Online Harms Deconstructed: the Initial Consultation Response. *Inforrm*. https://inforrm.org/2020/02/20/online-harms-deconstructed-the-initial-consultation-response-graham-smith

Taddeo, M. (2020). The Civic Role of OSPs in Mature Information Societies. In G. Frosio (Ed.), *Oxford Handbook of Online Intermediary Liability*. https://dx.doi.org/10.2139/ssrn.3584187

Tambini, D. (2019). The differentiated duty of care: a response to the Online Harms White Paper. *Journal of Media Law, 11*(1), 28-40. https://doi.org/10.1080/17577632.2019.1666488

The Guardian view on online harms: white paper, grey areas. *The Guardian*. https://www.theguardian.com/commentisfree/2019/apr/08/the-guardian-view-on-online-harms-white-paper-grey-areas

Theil, S. (2019). The Online Harms White Paper: comparing the UK and German approaches to regulation. *Journal of Media Law*, *11*(1), 41–51. https://doi.org/10.1080/17577632.2019.1666476

Tomlinson, H. (2019). Online Harms White Paper: Two comments on harms. *Inforrm*. https://inforrm.org/2019/07/05/online-harms-white-paper-two-comments-on-harms-hugh-tomlinson-qc/

Walker, C. and Akdeniz, Y. (1998). The governance of the Internet in Europe with special reference to illegal and harmful content. *Criminal Law Review*, 5-19.

Wragg, P. (2019). Tackling online harms: what good is regulation? *Communications Law, 24*(2), 49-51.

Regulations Cited

Chilean Regulations

Decree No. 12. (May 12, 2023). Creating a ministerial advising committee of the Ministry of Science, Technology, Knowledge, and Innovation with the name "Advisory Committee Against Disinformation" [Ministry of Science, Technology, Knowledge, and Innovation]. [REPEALED] https://bcn.cl/3e7zs

Decree No. 5. (March 6, 2024). Putting an end to "Advisory Committee Against Disinformation" [Ministry of Science, Technology, Knowledge, and Information]. https://bcn.cl/3khgd

Law No. 18314. (May 17, 1984). Identifying terrorist behavior and establishing punishments. https://bcn.cl/2k8cn

Law No. 19733. (June 4, 2001). On freedoms of opinion and information and practice of journalism. https://bcn.cl/2f8z0

Law No. 21522. (December 30, 2022) introducing a new paragraph under Title VII of Book II of the Criminal Code, relative to the commercial sexual exploitation and pornographic material of children and adolescents. https://bcn.cl/3ayhg

European Regulations

Council Framework Decision 2008/913/JHA. (November 2008, 2008). On combating certain forms and expressions of racism and xenophobia by means of criminal law http://data.europa.eu/eli/dec_framw/2008/913/oj

Final Communication of the European Commission COM/96/487. (October 16, 1996). Illegal and harmful content on the internet. https://eur-lex.europa.eu/procedure/EN/20878

Final Communication of the European Commission COM/2015/0192 (May 6, 2015). A

Digital Single Market Strategy for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52015DC0192

Final Communication of the European Commission COM/2016/288. (May 25, 2016). Online Platforms and the Digital Single Market: Opportunities and Challenges for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016DC0288

Final Communication of the European Commission COM/2017/0555. (September 28, 2017). Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017DC0555

Directive (EU) 2018/1808 of the European Parliament and of the Council. (November 14, 2018). Amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities. http://data.europa.eu/eli/dir/2018/1808/oj

Other Documents

Bill C-63: An Act to enact the Online Harms Act, to amend the Criminal Code, the Canadian Human Rights Act and An Act respecting the mandatory reporting of Internet child pornography by persons who provide an Internet service and to make consequential and related amendments to other Acts. (2021). 1st Reading, Feb. 26, 2024, 44th Parliament, 1st session. https://www.parl.ca/LegisInfo/en/bill/44-1/C-63

House of Representatives. (October 10, 2018). Gazette 12164-07. Amending the Criminal Code with the purpose of punishing the non-consensual dissemination of material with sexual connotations or of a sexual nature.