CENTER FOR DEMOCRACY & TECHNOLOGY

Positive Intent Protections:
Incorporating a *Good Samaritan principle* in the EU Digital Services Act

Authored by Joan Barata,
For the Center for Democracy & Technology

Executive Summary

The "Good Samaritan" principle ensures that online intermediaries are not penalized for good faith measures against illegal or other forms of inappropriate content. This is a rule that applies to concrete types of intermediaries, particularly those providing hosting services. When intermediaries are granted immunity for the content they handle, this principle in fact incentivizes the adoption and implementation of private policies regarding illegal and other types of lawful but offensive or undesirable content.

The principle finds one of its earliest and most acknowledged embodiments in Section 230(c) of the Communications Act of 1934 (as amended by the Telecommunications Act of 1996). Section 230 has played a fundamental role in the development of the Internet as we know it. Under the protections set by U.S. law, intermediaries have the incentive to operate and expand their businesses under a predictable legal regime, to moderate the content they share, and specifically to deal with certain forms of objectionable speech.

At the European level, the e-Commerce Directive (ECD) contains the general intermediary liability regime applicable to hosting services and establishes a series of provisions regarding the imposition of possible monitoring obligations to intermediaries. Intermediaries enjoy liability immunities inasmuch as they perform a role of a mere technical, automatic, and passive nature. This requirement of "passivity" is compatible with certain activities identified by the case law of the CJEU. However, intermediaries become liable in cases where they fail to act expeditiously to remove or to disable access to the illegal content upon obtaining knowledge or awareness, or they are simply proven to have overlooked a particular illegality when implementing voluntary and proactive monitoring measures in such a way as to create actual or constructive knowledge that strips them of immunity.

This legal framework, however, does not adequately promote the adoption of voluntary and proactive content moderation policies by private intermediaries, but rather the opposite. The more that intermediaries play an active role in monitoring the content they host, the more likely it becomes that they will find a potentially illegal piece of content. In this context the chances of overlooking a particular illegality, and therefore the risk of liability, grow significantly.

In order to incentivize content moderation under the Good Samaritan principle, and thereby enable intermediaries to address problematic but lawful content on their services, the paper sets a number of recommendations for the Digital Services Act (DSA). Given the importance of a strong liability framework to promote freedom of expression, access to information, and innovation online, the future DSA needs to keep the liability protections already present in

the ECD. At the same time, it also needs to create additional clarity about the scope and requirements in notice-and-action systems. In general, intermediaries should not be required to make determinations of illegality of third-party content; that is the function of courts. Uploaders of content should have the right to issue a counter-notice, and the framework should include penalties for notices sent in bad faith, among others. Exceptions to these general rules should be limited and narrowly defined.

Moreover, intermediaries should be transparent regarding the impact of their content moderation systems, and develop mechanisms to evaluate their effectiveness. Reporting mechanisms for content that is illegal and content as violating the service's own policies should be kept distinct, so that it is clear whether there is an allegation of illegality. Liability penalties should not arise from notifications of violations of content policies or Terms of Service. Should intermediaries be subject to duties aimed at preventing and tackling the dissemination of illegal content, these duties need to be commercially reasonable, transparent, proportionate, and generally flexible. Such obligations should not focus on the outcomes of content moderation processes, so as to avoid over-removal of lawful speech. Recognizing that there is no one-size-fits-all approach and maintaining flexibility for different content moderation practices can then enable effective Good Samaritan moderation of harmful, but not illegal, content.

CENTER FOR DEMOCRACY & TECHNOLOGY

I. Introduction

The so-called *Good Samaritan principle* ensures that online intermediaries are not penalized for taking steps to restrict illegal or other forms of inappropriate content.[1] This principle is particularly relevant for intermediaries providing hosting services, who tend to engage in more granular content moderation, but can apply to a relatively wide range of intermediaries that provide services for online storage, distribution, and sharing; social networking, collaborating and gaming; or searching and referencing[2]. The application of the *Good Samaritan principle* presents relevant implications vis-à-vis services provided by social media platforms like Facebook or Twitter, content sharing platforms such as YouTube or Vimeo, and search engines like Google or Yahoo, due to their role as facilitators of the exercise of the right to freedom of expression by users.

This rule is usually presented as protective of the activities and interests of intermediaries. As this paper will discuss, when intermediaries are granted immunity for the content they handle, the law is in fact incentivizing the adoption and implementation of private policies regarding illegal and other types of content that is lawful, but that may be offensive or undesirable in a given context. The content moderation systems that these intermediaries develop will reflect different interests, including the protection of the right to freedom of expression, the implementation of certain business models, and the avoidance of socially undesirable or harmful speech. Such interests are often intertwined and may present interesting reciprocal tensions.

The debate on regulation of content moderation systems contains a fundamental tension: On the one hand, States and certain civil society groups tend to ask intermediaries to make use of their own private regulatory tools to eradicate harmful and undesirable content, especially manifestations of hatred, disinformation, certain forms of propaganda, references to criminal acts, and other similar behaviours. On the other hand, organizations dedicated to the protection and promotion of freedom of expression, international human rights organizations, and even some governments have expressed their concern that global private companies, such as Google or Facebook, often restrict or simply eliminate ideas, opinions, and other content published by users on the basis of internal rules that are considered unjustified, abusive, and ambiguous. This tension also reveals different ways of understanding freedom of expression in a globalized world, even between States fitting under the category of liberal democracies.

[1] CDT, "Nine Principles for Future EU Policymaking on Intermediary Liability", 2019. Available online at: https://cdt.org/insights/nine-principles-for-future-eu-policymaking-on-intermediary-liability/.

[2] See the comprehensive and detailed categorization provided by Joris van Hoboken, João Pedro Quintais, Joost Poort, Nico van Eijk, "Hosting intermediary services and illegal content online. An analysis of the scope of article 14 ECD in light of developments in the online service landscape", European Commission - DG Communications Networks, Content & Technology and IViR, 2018. Available online at: https://op.europa.eu/en/publication-detail/-/publication/7779caca-2537-11e9-8d04-01aa75ed71a1/language-en.

II. The *Good Samaritan Principle* as Enshrined in the US Legal System

The principle finds one of its earliest and most explicit embodiments in Section 230(c) of the Communications Act of 1934 (as amended by the Telecommunications Act of 1996).

Section 230(c), titled "Protection for 'Good Samaritan' blocking and screening of offensive material", contains two different immunities regarding the use or provision of "interactive computer services", enshrined under (c)(1) and (c)(2)(A) respectively. The first immunity establishes that no user or provider "shall be treated as the publisher or speaker of any information provided by another information content provider". The second main immunity shields the same subjects from being held liable on account of:

> "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected (...)"

Despite the fact that all of the immunities contained in Section 230(c) are labelled as *Good Samaritan* protections, the most common use of this term in American scholarship and jurisprudence focuses on the provisions included in (c)(2)(A). This is also the approach taken by this paper. In other words, I will use "Good Samaritan" to mean immunities for *taking content down*, as distinct from immunities for *leaving content up.* (Also note that both immunities have a series of exceptions mentioned in 230(e), regarding criminal law, intellectual property law, State law, privacy law, and, as of 2018, sex trafficking law.)

The adoption of Section 230 in 1996 was preceded by a few court decisions that generated a significant level of legal uncertainty regarding the circumstances under which an intermediary could be held liable for third-party content, particularly in cases where these intermediaries play a role in controlling or moderating the third-party content[3]. The *Good Samaritan principle* of Section 230 aims at guiding judicial decisions in these areas, as well as providing a clear legal framework for intermediary companies to operate. It is also important to note that while the First Amendment of the Constitution itself would shield intermediaries from some types of liability for content provided by third parties, it may not require immunizing intermediaries for editing or deleting user content[4]. Therefore,

[3] In particular, *Cubby Inc. v. CompuServe Inc.* (https://law.justia.com/cases/federal/district-courts/FSupp/776/135/2340509/) and *Stratton Oakmont, Inc. v. Prodigy Services Co.* (https://h2o.law.harvard.edu/cases/4540).

[4] Prominent Yale freedom of expression scholar Jack Balkin explains that "Some aspects of intermediary immunity are probably required by the Constitution, so that if Congress repealed § 230, certain constitutional protections would still be in force. For example, it might be unconstitutional to hold digital curators strictly liable for any defamatory or obscene content that appears on their sites. But the boundaries of constitutional protection are uncertain. Would a negligence standard be sufficient? What about other kinds of unlawful content?", in "Free Speech is a Triangle", *Columbia Law Review* Vol. 118: 2011 (2018). An interesting and nuanced approach is also provided by Jeff Kosseff in "First amendment protection for small platforms", *Computer Law & Security Review 35* (2019) 199–213.

protections contained in Section 230, and particularly (c)(2)(A), clearly increase, beyond the First Amendment, predictability in this area and reduce adjudication costs.[5]

The text of 230(c)(2)(A) includes references to "actions taken in good faith" against a non-exhaustive list of "objectionable" categories of content. Recent debates about Section 230 have included arguments that these references significantly limit the scope of 230(c)(2), or even (c)(1)'s, liability shield and impose a kind of "neutrality" obligation on intermediaries, but these arguments find little basis in either the text or intent of Congress in passing the statute, or in the two decades of case law interpreting it. The element of "actions taken in good faith" has not been enforced as strictly opposite to malicious or capricious behaviour, nor in the sense of requiring any specific or articulated proof of such *bona fides*, but essentially as demanding from the intermediary's side some plausibility or a minimal amount of justification.[6] The utility of the "good faith" element continues to be debated, with legal scholars, including Eric Goldman, arguing that it only invites judicial confusion and increases the chances that both parties will incur more adjudication costs only to reach the same result: a prevailing defendant[7]. The scholar Annemarie Bridy, on the other hand, argues that U.S. courts should interpret "good faith" in Section 230 to require clarity and consistency in enforcement, thus providing an avenue of appeal for users who believe their content has been wrongly removed[8].

As for the second requirement, Section 230 refers to a series of broad categories of content in order to justify and protect moderation decisions taken by intermediaries ("obscene, lewd, lascivious, filthy, excessively violent, harassing"). The most important (and subject to interpretation) element is the reference to content "otherwise objectionable". Courts have generally seen this criterion to be a catchall notion to cover any type of content that intermediaries themselves consider, under their own criteria and internal standards, to be objectionable. It is also important to note that constitutional protection of the underlying speech is not a factor that intermediaries need to take into account when deciding to take actions vis-à-vis content under the immunities provided by Section 230. They are explicitly

[5] See Eric Goldman, "Why Section 230 Is Better than the First Amendment", 2 *Notre Dame Law Review Reflection* 34 (2019).

[6] Occasionally, a court will revisit the implications of the "good faith" element, but the most recent example of this, the Ninth Circuit's opinion *Enigma v. Malwarebytes*, also confuses the matter further by reading the "good faith" element of 230(c)(2)(A) into the liability protections afforded by 230(c)(2)(B) to makers of filtering and blocking tools. *See* Eric Goldman, "Ninth Circuit Doubles Down on Bad Ruling That Undermines Cybersecurity–Enigma v. Malwarebytes", *Technology & Marketing Law Blog*, 9 January 2020. Available online at:

https://blog.ericgoldman.org/archives/2020/01/ninth-circuit-doubles-down-on-bad-ruling-that-undermines-cybersecurity-enigma-v-malwarebytes.htm Malwarebytes is seeking review of this opinion by the Supreme Court of the United States.

[7] Eric Goldman, "Online User Account Termination and 47 U.S.C. § 230(c)(2)", 2 *UC Irvine Law Review* 659 (2012). Goldman also notes that many courts have avoided the question of the "good faith" element in 230(c)(2)(A) by simply interpreting 230(c)(1)'s protections against publisher liability to include a liability shield for content moderation and other "editorial" activity.

[8] Annemarie Bridy, "Remediating Social Media: A Layer-Conscious Approach," 24 *B.U. J. Sci. & Tech. L.* 193, 221-222 (2018).

shielded in their moderation practices "whether or not such material is constitutionally protected".

The main conclusions and outcomes derived from the existence of the *Good Samaritan principle* in U.S. law (putting aside legal claims that are simply not covered by Section 230, such as violations of copyright or federal criminal law) are therefore the following:

a. Intermediaries are not liable for the third-party content that they share or decide, in any circumstance, to keep available. In particular, this means that content hosts are not liable for illegal content that they fail to detect or assess.
b. Intermediaries are free to set their own content policies, which may essentially be tailored to the characteristics of their users and their commercial, social, or even political (should they have them) interests. This means that, as a matter of principle, there is no applicable legal requirement regarding the content and nature of such policies. Moreover, inasmuch as service providers are private actors, they can also proscribe constitutionally protected speech.
c. Intermediaries are, in particular, not liable for content moderation decisions taken on the basis of the policies mentioned above. This means that service providers are actually encouraged to ban, police, and remove not only presumed illegal posts, but also lawful, yet still harmful or offensive content. This also enables providers to invite their users to flag inappropriate content, without a fear that such notifications will create a liability risk for the provider. As previously mentioned, decisions taken in this area are, as a matter of principle, shielded from third-party liability claims.

Section 230 has played a fundamental role in the development of the Internet as we know it. Under the protections set by U.S. law, intermediaries have the incentive to operate and expand their businesses under a predictable legal regime, to moderate the content they share, and specifically to deal with certain forms of *objectionable* or harmful speech. In addition to this, Section 230 not only respects and further develops the free speech protections enshrined in the First Amendment, but also creates a framework which prevents pressure on intermediaries to implement over-broad or "preventive" restrictions regarding content alleged or suspected to be illegal, as would be the case if broad liability provisions were in place. (This dynamic is also known as collateral censorship, or erring on the side of caution.)

The Section 230 framework does not answer every question relating to the role of different information intermediaries in our societies. Section 230 shields intermediaries from pressure to take down speech (either legal or illegal), but it is undeniable that content moderation decisions present a relevant impact on the individual exercise of the right to freedom of expression. The fact that intermediaries provide an openly accessible space to speak, neither deprives this space from its intrinsically private nature, nor, under US law, creates an obligation for intermediaries to protect users' right to freedom of expression. Yet at the same time, as scholar Daphne Keller has outlined, major platforms can restrict our speech more effectively than any government in history. And these restrictions are not applied necessarily on the basis of human, balanced decisions, but using automated tools

that can take swift decisions on sensitive content. The same determinations might take courts months or years[9].

Despite the absence of sufficient transparency, there are indications that big platforms already have a long record of mistaken or harmful decisions in this area, inside and outside the United States[10]. In a country where platforms cannot be treated as State actors, mechanisms such as market forces, competition, and at last instance antitrust law, are commonly presented as the checks and balances of companies' behaviour[11]. In addition to potential legal remedies, there are interesting proposals coming, particularly from civil society and academia, aimed at establishing certain self-regulatory codes or principles that would improve transparency and accountability of platforms' decisions, including initiatives such as the Santa Clara Principles on Transparency and Accountability in Content Moderation[12].

III. Positive Intent Protections and EU Legislation: the e-Commerce Directive

Article 14 of the e-Commerce Directive (ECD)[13] contains the general intermediary liability regime applicable to hosting services at the EU level:

> "1. Where an information society service is provided that consists of the storage of information provided by a recipient of the service, Member States shall ensure that the service provider is not liable for the information stored at the request of a recipient of the service, on condition that:
>
> (a) the provider does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent; or
>
> (b) the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information.

[9] Daphne Keller, "Facebook Restricts Speech by Popular Demand", *The Atlantic*, 22 September 2019. Available online at: https://www.theatlantic.com/ideas/archive/2019/09/facebook-restricts-free-speech-popular-demand/598462/

[10] See Jillian C. York, Karen Gullo, "Offline/Online Project Highlights How the Oppression Marginalized Communities Face in the Real World Follows Them Online", *Electronic Frontier Foundation*, 6 March 2018 (https://www.eff.org/deeplinks/2018/03/offlineonline-project-highlights-how-oppression-marginalized-communities-face-real), Billy Perrigo, "These Tech Companies Managed to Eradicate ISIS Content. But They're Also Erasing Crucial Evidence of War Crimes, *Time*, 11 April 2020 (https://time.com/5798001/facebook-youtube-algorithms-extremism/?xid=tcoshare), and "When Content Moderation Hurts", *Mozilla*, 4 May 2020 (https://foundation.mozilla.org/en/blog/when-content-moderation-hurts/).

[11] See the comprehensive analysis of these matters by Daphne Keller, "Who Do You Sue? State and Platform Hybrid Power over Online Speech", *Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper* No. 1902, 29 January 2019. Available online at: https://www.lawfareblog.com/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech.

[12] Available online at: https://santaclaraprinciples.org.

[13] 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.

> 2. Paragraph 1 shall not apply when the recipient of the service is acting under the authority or the control of the provider.
> 3. This Article shall not affect the possibility for a court or administrative authority, in accordance with Member States' legal systems, of requiring the service provider to terminate or prevent an infringement, nor does it affect the possibility for Member States of establishing procedures governing the removal or disabling of access to information."

Article 15 also establishes a series of provisions regarding the imposition of possible monitoring obligations to intermediaries:

> "1. Member States shall not impose a general obligation on providers, when providing the services covered by Articles 12, 13 and 14, to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity.
> 2. Member States may establish obligations for information society service providers promptly to inform the competent public authorities of alleged illegal activities undertaken or information provided by recipients of their service or obligations to communicate to the competent authorities, at their request, information enabling the identification of recipients of their service with whom they have storage agreements."

The higher complexity and detail of European rules makes it somewhat more difficult to describe to what extent a *Good Samaritan principle* is in place and applied in Europe. On the basis of the provisions mentioned above, hosting platforms (as part of the broader category of information society service providers) in Europe are not liable for, or in connection to, content moderation decisions on the basis of several requirements and conditions:

Knowledge and/or Awareness

In order to retain immunity, platforms must not have actual knowledge of illegal activity or information (for criminal law claims), and must not be aware of facts or circumstances from which the illegal activity or information is apparent (for civil law claims). According to the interpretation provided by the Court of Justice of the European Union (CJEU) in the *L'Oréal* case[14], rules set out in article 14.1.a) of the ECD "must be interpreted as covering every situation in which the provider concerned becomes aware, in one way or another, of such facts or circumstances". In particular, these rules include situations where the intermediary achieves actual or specific (not presumed or constructed) knowledge of the illegality "as the result of an investigation undertaken on its own initiative", or receives a proper notification that allows the intermediary to become "actually aware of facts or circumstances on the basis of which a diligent economic operator should have identified the illegality". This is of course without prejudice, according to article 14.3, to "the possibility for a court or administrative authority, in accordance with Member States' legal systems, of requiring the service provider to terminate or prevent an infringement".

[14] Judgement of 12 July 2011, case C-324/09.

It is important to underscore that knowledge or awareness do not equal the need to act upon any kind of notice in order to avoid liability under article 14. Recital 46 suggests that intermediaries need to take proper and balanced decisions in this area, bearing particularly in mind the "observance of the principle of freedom of expression and of procedures established for this purpose at national level". In addition to this, the CJEU has specified, also in the *L'Oréal* case, that notices need to be precise and substantiated.

Expeditious Removal

As it has been shown, article 14.1.b) of the ECD requires, for the liability exemption to be applied, that once knowledge or awareness are acquired, the intermediary acts "expeditiously". There is no particular development or a more specific definition, at least at the level of EU legislation, of what is this expeditiousness, which will basically depend on the resources and capacities of the intermediary, as well as the way in which knowledge or awareness is obtained.

Neutral Provision of the Service

The current CJEU case law uses an additional controversial standard to determine the application of liability immunities to hosting intermediaries. This approach is based on the wording of Recital 42 of the ECD, which provides that the liability exemptions are applicable when the role of the intermediary "is of a mere technical, automatic and passive nature, which implies that the information society service provider has neither knowledge of nor control over the information which is transmitted or stored". Despite the fact that a consistent reading of this Recital would suggest that this neutrality requirement would only be applicable vis-à-vis "mere conduit" and "caching" activities, and thus to the immunities established in articles 12 and 13 of the Directive, the CJEU has also considered it applicable to hosting activities.

In the *Google France* ruling[15], and regarding the web search and advertising services provided by this company, the CJEU states, as per its "technical, automatic and passive" nature, that "the mere facts that the referencing service is subject to payment, that Google sets the payment terms or that it provides general information to its clients cannot have the effect of depriving Google of the exemptions from liability". Equally, the decision also affirms that "concordance between the keyword selected and the search term entered by an internet user is not sufficient of itself to justify the view that Google has knowledge of, or control over, the data entered into its system by advertisers and stored in memory on its server". In the *L'Oréal* case, the Court limits liability to cases where the intermediary "plays an active role of such a kind as to give it knowledge of, or control" over the hosted content. It would not be considered an active role when "the operator of an online marketplace stores offers for sale on its server, sets the terms of its service, is remunerated for that service and provides general information to its customers". However, it does qualify as an active role to provide "assistance which entails, in particular, optimising the presentation of the offers for sale in question or promoting those offers". This being said, there are still some pending cases before the CJEU where the Court will have the opportunity to provide

[15] Judgment of 23 March 2010, joined cases C-236/08 to C-238/08.

some additional clarifications[16]. This is an important question to clarify, as thus far it is difficult to determine the general principles according to which intermediaries' interventions can clearly be classified as active or passive (with the corresponding consequences in terms of liability), as we only have a few specific examples derived from individual court cases.

These overlapping elements of the ECD's liability framework indicate that *Good Samaritan* protections may be applicable to intermediaries under EU legal provisions and the case law of the CJEU, although they are built in a more complex way than in the U.S. Besides these concrete examples there are still plenty of other possible interventions or activities, particularly regarding content moderation practices, that remain in a grey area.

To get a more detailed picture of these problems of interpretation it is now necessary to analyse the provisions regarding the non-imposition of general monitoring obligations, also included in the Directive.

No General Obligation to Monitor

The first paragraph of Article 15 prohibits the imposition of general content monitoring obligations, as well as obligations "to (actively) seek facts or circumstances indicating illegal activity". Recital 47 clarifies that these prohibitions do not preclude "monitoring obligations in a specific case". In the words of the CJEU in the *L'Oréal* decision, "the measures required of the online service provider concerned cannot consist in an active monitoring of all the data of each of its customers", although it can be ordered to take specific measures in order to terminate a particular infringement or facilitate the identification of an individual offender. In *Scarlet Extended*[17] and *SABAM*[18], the Court has specifically established that national courts are precluded from issuing injunctions against hosting service providers which require them to install a system for filtering, when such a system would actively monitor all the data of each of their customers in order to prevent future legal infringements. Recital 48 also provides some guidance in this area by generally stating that the restrictions included in Article 15 do not "affect the possibility for Member States of requiring service providers, who host information provided by recipients of their service, to apply duties of care, which can reasonably be expected from them and which are specified by national law, in order to detect and prevent certain types of illegal activities."

It is important to note that, in light of this jurisprudence, Articles 14 and 15 would in principle need to be read and interpreted in a separate manner. While the former establishes knowledge and awareness thresholds and parameters in order for hosting intermediaries to keep their immunities, the latter frames the possible imposition of specific and targeted content monitoring duties. In this second case, legal responsibilities may be

[16] See for example the case of LF v Google LLC, YouTube Inc., YouTube LLC, Google Germany GmbH, pending case available online at: http://curia.europa.eu/juris/document/document.jsf;jsessionid=61CBAAD992CF5936228BF1C639B19FC5?text=&docid=211267&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1020534.

[17] Judgment of 24 November 2011, case C-70/10.

[18] Judgement of 16 February 2012, case C-360/10.

demanded according to national legislation, only when such duties have not been properly fulfilled.

In the recent case of *Eva Glawischnig-Piesczek*[19], the CJEU and the Advocate General (AG) Szpunar seem to endorse an interpretation of Articles 14 and 15 that finds that specific monitoring obligations are valid and possible only when they do not put intermediaries in the position of becoming liable under the parameters of Article 14. This may lead to the conclusion that these obligations are acceptable inasmuch as they can be fulfilled with recourse to "passive" automated search tools and technologies. Conversely, and according to the AG, applying human error-correction protocols to fulfil the obligation of finding content "identical" to the previously identified as illegal may make the platform (Facebook, in this case) lose its immunity in terms of Article 14[20].

Some recently adopted legislation, in particular the Copyright Directive[21], as well as legislative proposals under discussion in the field of terrorist content online[22], seem to incorporate new obligations regarding the adoption of proactive measures. Such measures would require intermediaries to detect, identify, remove, or disable access, and even prevent the re-uploading of certain pieces or types of content. In some cases, including in the proposed legislation regarding terrorist content, these measures can only be seen as an actual (and declared) derogation of the non-monitoring principle enshrined in the e-Commerce Directive.[23]

Proactive Measures

Intermediaries are also allowed to voluntarily adopt their own content moderation and monitoring rules and enforce them. This behaviour is even promoted by several initiatives of EU institutions, including the European Commission's Code of Conduct for "countering illegal speech online", which was launched with Facebook, Microsoft, Twitter, and YouTube

[19] Judgment of 3 October 2019, case C-18/18.

[20] See the consequences that the extended use of this criterion would have in terms of human rights impact in Daphne Keller "Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion", *Stanford CIS White Paper*, 2019. Available online at: http://cyberlaw.stanford.edu/blog/2019/09/filtering-facebook-introducing-dolphins-net-new-stanford-cis-white-paper-or-why See also Daphne Keller, "Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling", GRUR International, ikaa047, 2020. Available online at: https://academic.oup.com/grurint/advance-article-abstract/doi/10.1093/grurint/ikaa047/5831378?redirectedFrom=fulltext.

[21] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

[22] Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. Documents available here: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52018PC0640.

[23] See a more detailed description of these provisions and a critical approach in Aleksandra Kuczerawy "To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive", *Balkinization*, 2019. Available online at: https://balkin.blogspot.com/2019/05/to-monitor-or-not-to-monitor-uncertain.html and Joan Barata Mir "New EU proposal on the prevention of terrorist content online: an important mutation of the e-commerce intermediaries' regime", *Stanford CIS White Paper*, 2018. Available online at: http://cyberlaw.stanford.edu/publications/new-eu-proposal-prevention-terrorist-content-online-important-mutation-e-commerce.

in May 2016.[24] Other relevant initiatives are the Communication of the European Commission on tackling illegal content online of 2017[25] and the companion Recommendation of 2018.[26] In particular, the Communication stresses the following:

> "Online platforms should, in light of their central role and capabilities and their associated responsibilities, adopt effective proactive measures to detect and remove illegal content online and not only limit themselves to reacting to notices which they receive. Moreover, for certain categories of illegal content, it may not be possible to fully achieve the aim of reducing the risk of serious harm without platforms taking such proactive measures.
>
> The Commission considers that taking such voluntary, proactive measures does not automatically lead to the online platform losing the benefit of the liability exemption provided for in Article 14 of the E-Commerce Directive."

This text seems to suggest that online intermediaries are allowed to adopt *Good Samaritan* measures and still be protected under the immunities provided in Article 14 of the Directive. These measures can be of "proactive" nature (thus not limited to third-party notices) and directed at detecting and properly identifying illegal content online. However, the Communication also emphasizes that the same Article 14 results in the obligation for platforms to "act expeditiously to remove or to disable access to the information in question upon obtaining such knowledge or awareness". Only when such expeditious reaction takes place are intermediaries able to keep the mentioned immunity.

Contrary to the U.S.' Section 230, this legal framework does *not* promote the adoption of voluntary and proactive content moderation policies by private intermediaries, but rather the opposite. The more that intermediaries play an active role in monitoring the content they host, the more likely it becomes that they will find a potentially illegal piece of content which would at least require some cautious consideration. Moreover, in this context the chances of overlooking a particular illegality, and therefore the risk of liability, grow significantly[27]. This tension was a feature of the ruling of the European Court of Human Rights in the landmark *Delfi v. Estonia* case, which held the online publication Delfi responsible for the hateful comments posted by readers in reaction to an article[28]. The ECHR

[24] Available online at: https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300.
[25] Communication of the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions on Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms. COMM (2017) 555 final. 28 September 2017.
[26] Commission Recommendation of 1 March 2018 on measures to effectively tackle illegal content online. C (2018) 1177 final.
[27] Aleksandra Kuczerawy, "The EU Commission on voluntary monitoring: Good Samaritan 2.0 or Good Samaritan 0.5?", *KU Leuven CiTiC* 2019. Available online at: https://www.law.kuleuven.be/citip/blog/the-eu-commission-on-voluntary-monitoring-good-samaritan-2-0-or-good-samaritan-0-5/.
[28] Case of *Delfi v. Estonia*. Judgement of 15 June 2015. Application no. 64569/09. See Joan Barata Mir and Marco Bassini, "Freedom of expression in the Internet: main trends of the case law of the European Court of Human Rights", in Pollicino, O., *Internet Law and Constitutional Adjudication*, London: Routledge 2015.

has since developed and clarified this doctrine, particularly regarding situations of platform liability not involving hate speech or calls to violence in a series of posterior rulings[29].

Thus, all together EU law currently encompasses an extremely limited and vague version of the *Good Samaritan principle*, based on the following elements:

a. Intermediaries enjoy liability immunities inasmuch as they perform a role of a mere technical, automatic, and passive nature. This requirement of "passivity" is compatible with certain activities identified by the case law of the CJEU. In any case, as Sartor has explained, making the protection conditional on passivity would induce a hands-off approach that results both in an increased quantity of online illegalities, and in the failure to satisfy the users who prefer not to be exposed to objectionable or irrelevant material[30].
b. Voluntary proactive measures to monitor, detect, and remove illegal content online do not necessarily lead the online platform in question to lose the benefit of the liability exemption.
c. However, intermediaries become liable in cases where they fail to act expeditiously to remove or to disable access to the illegal content upon obtaining knowledge or awareness, or they are simply proven to have overlooked a particular illegality when implementing voluntary and proactive monitoring measures in such a way as to create actual or constructive knowledge that strips them of immunity.

IV. Towards a Strong *Good Samaritan Principle* in the EU Legal System

The European Commission has committed to submit a proposal for a Digital Services Act (DSA) legislative package which, among other things, will aim to revise the provisions contained in the e-Commerce Directive. This intention has already triggered the elaboration of several opinions and reports, particularly from different committees of the European Parliament.[31]

One of the main issues at stake within the framework of the mentioned process is whether liability exemption provisions contained in the ECD must be kept, reduced, or improved.

[29] See Dick Voohoof, "The Court's subtle approach of online media platforms' liability for user-generated content since the 'Delfi Oracle'", *Strasbourg Observers* 10 April 2020. Available online at: https://strasbourgobservers.com/2020/04/10/the-courts-subtle-approach-of-online-media-platforms-liability-for-user-generated-content-since-the-delfi-oracle/.

[30] Giovanni Sartor, "Providers Liability: From the eCommerce Directive to the future", in-depth analysis for the IMCO Committee commissioned by the Policy Department for economic and scientific policy, Directorate-General for Internal Policies, European Parliament, 2017. Available online at: https://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf.

[31]Committee on the Internal Market and Consumer ProtectionCommittee on the Internal Market and Consumer Protection, Draft Report available at https://www.europarl.europa.eu/doceo/document/IMCO-PR-648474_EN.pdf; Committee on Legal Affairs, Draft Report available at https://www.europarl.europa.eu/doceo/document/JURI-PR-650529_EN.pdf; Committee on Civil Liberties, Justice, and Home Affairs, Draft Report: https://www.europarl.europa.eu/doceo/document/LIBE-PR-650509_EN.pdf.

Given the importance of a strong liability framework to promote freedom of expression, access to information, and innovation online, the future DSA needs to keep the liability protections already present in the ECD. At the same time, it should introduce new provisions aimed at further encouraging hosting service providers to moderate content in a proportionate manner and to devote appropriate efforts to tackle illegal and other forms of socially undesired content, without unnecessarily burdening users' right to freedom of expression. The resulting framework in the DSA should follow these principles:

Preserve Strong and Clear Baseline Liability Framework

A clear immunity from liability for infrastructure intermediaries—including those that provide "mere conduit" and "neutral hosting" services under the ECD framework—should be maintained. Such services should only ever face liability for third-party content for failure to remove or block specific content following a court order. It must be clear that intermediaries of all types do not have an obligation to actively monitor and identify illegal content, and that a failure to proactively identify illegal content does not make them become liable.

Create Clarity and Include Safeguards in Notice-and-Action Systems

For intermediaries that do engage in more granular moderation and curation of third-party content—"active hosts" in the ECD framework—the DSA should provide significant additional clarity about the scope and requirements of notice-and-action procedures. Intermediaries should not be required to make determinations of illegality of third-party content; that is only the function of courts. This means that intermediaries should not face liability for failing to remove content on the basis of a notice from actors other than courts, and should be able to challenge the validity of notices received from other government officials or private parties. In general, notices of alleged illegal content must include the identity of the issuing officer or private claimant, a citation to the specific legal authority or a clear explanation of the violation of the law, and the specific URL of the challenged content. The notice-and-action framework should also enable the individuals who posted the challenged content to provide a counter-claim or counter-notice to rebut the claim of illegality of their speech, and the framework should include penalties for notices sent in bad faith. Crucially, the question of notice-and-action-based liability for illegal content should not be conflated with intermediaries' general content moderation practices. Legislation should focus on the principle that content creators are responsible, under the law, for their online speech and behaviour.

Structural or Systemic Oversight Must Not Disincentivize *Good Samaritan* Moderation

If intermediaries are subject to certain structural or systemic[32] duties aimed at preventing and tackling the dissemination of illegal content, these duties need to be commercially reasonable, transparent, proportionate, and generally flexible. Such obligations should not focus on the outcomes of content moderation processes, i.e.

[32] This term is used in a similar meaning by Daphne Keller in "Systemic Duties of Care and Intermediary Liability", *CIS Blog*, 2020. Available online at: https://cyberlaw.stanford.edu/blog/2020/05/systemic-duties-care-and-intermediary-liability.

intermediaries should not be evaluated on whether they have removed "enough" illegal content, as this creates a strong incentive towards over-removal of lawful speech. Intermediaries should not face penalties, for example, for failing to "consistently" or "comprehensively" enforce their policies against illegal content, as this creates a disincentive towards having specific and nuanced policies aimed at combating abuse of their platforms. Legal regimes must clearly differentiate administrative responsibility related to failure to fulfil regulatory obligations from loss of immunity regarding hosted content. Sanctions should only be applied in cases of demonstrated systemic failure to respond to valid notifications of illegal content.

Commitments to Transparency and Accountability in Content Moderation

The legal framework should encourage intermediaries to be transparent regarding the impact of their content moderation systems, and to develop mechanisms to evaluate their effectiveness. Intermediaries should have in place adequate, accessible, and easy-to-use mechanisms to report illegal content and to flag content as violating the service's own policies, though these reporting mechanisms should be kept distinct so that it is clear whether there is an allegation of illegality against a particular post. No liability or penalties should arise from notifications of apparent violations of content policies or Terms of Service.

Maintain Flexibility for Different Approaches to Content Moderation

Effective content moderation will consist of different policies and practices for different types of services and different user-bases and communities. According to the standards set by the Council of Europe, and as a general principle applicable to the previous considerations, "States should take into account the substantial differences in size, nature, function and organisational structure of intermediaries when devising, interpreting and applying the legislative framework in order to prevent possible discriminatory effects." Enabling effective *Good Samaritan* moderation of harmful, but not illegal, content, requires recognizing that there is no one-size-fits-all approach, and ensuring that the legislative framework is not overly prescriptive as to the substance or method of content moderation, and does not create legal risk or onerous regulatory obligations that will discourage or constrain intermediaries' content moderation efforts.