REVISTA CHILENA
DE DERECHO Y TECNOLOGÍA

VOL. 13 (2024) • e70729 • DOI 10.5354/0719-2584.2024.70729
RECIBIDO: 20/5/2023 • APROBADO: 8/3/2024 • PUBLICADO: 7/4/2024

DOCTRINA

Anonymity and the challenges of regulating online harmful conducts

El anonimato y los desafíos de regular las conductas dañinas en línea

María Francisca Ossa Monge
Pontificia Universidad Católica de Chile

ABSTRACT The aim of this essay is to analyse the conflictive relationship between online anonymity and online harmful conducts that fall outside the scope of existing regulation. With this objective in mind, this essay analyses the increasing recognition of online harmful practices as a problem that needs to be tackled, and how anonymity, while virtuous in keeping the privacy of users to the best of its ability, can also enable online harmful conduct by providing a sense of a lack of accountability. Moreover, we explore the connections between the notions of freedom of speech and freedom of expression with anonymity and online harmful practices, as well as the challenges to regulate the latter with a limitation of the former. Finally, we take a glimpse at Lessig's regulation modalities (law, social norms, and architecture or "code"), and examples of how each of them encapsulate online harm and if measures regarding anonymity should be added to said modalities, in order to properly face online harmful practices as an increasing problem for the current and new generations with the ever-expanding presence of internet worldwide.

KEYWORDS Online harm, anonymity, regulation.

RESUMEN El objetivo de este ensayo es analizar la relación conflictiva entre el anonimato en línea y las conductas abusivas o dañinas que no caben dentro del marco regulatorio actual. Con este objetivo en la mira, este ensayo analiza el creciente reconocimiento de las prácticas dañinas que ocurren en línea como un problema al que hay que enfrentarse, y cómo el anonimato, siendo útil para mantener la privacidad de sus usuarios, también puede propiciar la comisión de prácticas dañinas al proporcionar a sus usuarios falta de responsabilización. Asimismo, se analizarán las conexiones entre las nociones de libertad de palabra y de expresión con el anonimato y las prácticas dañinas en línea, así como los desafíos que presenta la regulación de estas mediante la limitación de aquello. Finalmente, se da una mirada a las modalidades de regulación de Lessig (ley, normas sociales y arquitectura o «código»), ejemplos de cómo cada una

de ellas tratan las conductas dañinas que ocurren en línea, y si medidas con respecto al anonimato debieran añadirse a dichas modalidades de regulación para enfrentar a las conductas dañinas que ocurren en línea. Esto último es un creciente problema al que se enfrentan las generaciones contemporáneas y venideras con la presencia del internet que está siempre en expansión alrededor del mundo.

PALABRAS CLAVE Contenido dañino en línea, anonimato, regulación.

Introduction

Human nature, in all its complexity, is not likely to change, not even due to the increasing use and abuse of online interaction and functioning in cyberspace. Therefore, if there is room for harmful conduct, such as hate speech, deceit, and bullying in the real or offline world, it is not surprising that it also takes place online, which seems to have been the case since the very invention of the Internet (Franks, 2019: 137).

What seemed to be an enabling factor for online harmful practices, was rather an absence of an effective regulation of cyberspace, which for many years remained feeble to non-existent. If there was a degree of regulation, it was very distanced from that of the same practices in the real world, as if there was a dissociation between the person who incurs in them behind a screen and the person who incurs in them in plain sight, as in the middle of a public park. This lack of effective regulation has also led to other issues, namely digital vigilantism by users who intervene when the explicit or tacit rules of the site or platform are broken, sometimes with no sense of limits and incurring in vindicative harmful practices. This disassociation between these user-persona and their offline versions was —and to some extent still is— explained by the irreconcilable differences between the online and offline realities that make these two "spaces" extremely challenging to regulate in the same way.

We argue that online anonymity is one of those irreconcilable differences, and when combined with a lack of effective regulation of cyberspace, it may act as an enabler factor for harmful conduct, both *ex-ante* when disinhibiting the infringers and *ex-post* when concealing their identities, and therefore impeding the necessary sense of accountability regarding online actions.

This essay intends to analyse the link between online anonymity and harmful conduct on the Internet, the tensions between the attempts to limit said anonymity with the ideas of privacy and freedom of speech online, and finally an insight from a regulatory perspective following Lawrence Lessig's modalities of cyberspace regulation, to properly assess whether limiting online anonymity would be an efficient solution to the problem of online harm, or if there are some more eclectic, intermediate solutions.

Definitions and analysis framework

Online anonymity and pseudonymity

Online anonymity, in general, is an equivocal concept. For the purposes of this essay and the underlying discussions regarding anonymity's relation with online harm, we will follow Moore (2018: 169), who distinguishes between three different dimensions of online anonymity, namely: i) traceability,[1] which refers to "the extent to which your contributions can be traced to your real identity"; ii) durability, which refers to "the ease or difficulty with which online identities can be acquired and changed"; and iii) connectedness, which he characterizes as the "bridging across different platforms and contexts". This last dimension is strongly linked to the reputation of the user, considered on a global level rather than locally in a single platform (Moore, 2018: 169).

As we will see through this essay, analysing these different features or dimensions of anonymity is a useful exercise when assessing the tensions it can create regarding the privacy of users (linked to the traceability dimension of anonymity) or the lack of accountability for infringers —this could be linked to the dimensions of the durability of identities if, for instance, the platforms were to close the infringer account, and to the dimension of connectedness if other users were to dox[2] an infringer, hence jeopardizing the reputation of an anonymous or pseudonymous account.

As for pseudonymity, Véliz analyses this concept, which, "derived from the Greek 'false name', […] involves the identification of an author through a tag that does not correspond to her real name" (Véliz, 2019: 633). Throughout this essay, although acknowledging their different levels of influence in the different topics of our discussion (such as accountability and freedom of speech), we will refer to anonymity and pseudonymity indistinctly, as both refer to a lack of real-name engagement online, unless otherwise specified.

Online harmful conducts: A non-exhaustive concept

When it comes to online harm or online harmful practices, the definition in the literature on the topic tends to be elusive. Most authors simply mention the practices that can be categorised as "harmful" in the virtual space of the Internet. This is not surprising since there are many subjectivities, ethical boundaries, and dilemmas that are likely to be trespassed if, for instance, the legislator of a certain jurisdiction attempted to define such a concept (especially when talking about online harm that is not necessarily illegal). In this sense, we share Bernal's (2019) concerns regarding

1. For a detailed analysis of the dimension of traceability in anonymity, see Kaminski (2012: 815-896).

2. Term that indicates to publish private information about someone on the internet, without their permission and in a way that reveals their name, where they live, etcetera.

the UK's Online Harms White Paper's categorization of online harms, including both legal and illegal practices:[3]

> That the White Paper starts by referring to illegal and "unacceptable" content and activity should be a concern from the start. If something is really "unacceptable", it should be made illegal —and unless it is illegal, it should not be deemed unacceptable. If acceptability can be determined by policy or politics rather than the law, the scope for abuse, uncertainty and bias is enormous. Setting what amounts to a "moral" or "ethical" view of acceptability is a very slippery slope.[4]

Preliminarily speaking, we agree with Bernal's approach on the difficulties that come from considering some practices as "unacceptable" instead of straight-forwardly illegal. Nonetheless, we believe that for the time being, while there are harmful conducts that are not yet considered illegal and until they do become illegal or somehow cease to exist, analysing them and their potential link to anonymity remains relevant, especially considering the damage those practices can have on internet users or addressees of harmful conduct.

Therefore, for the purposes of this essay, the scope will be placed on the segment of harmful practices that are not yet illegal in a certain jurisdiction, that, as the Online Harms White Paper treats them, "online harms with a less clear definition": practices like cyberbullying, *trolling*, extremist content and activity, intimidation, disinformation, violent content, flaring, doxing, etcetera. These take place on social media platforms, public discussion forums, and platforms that allow users to share content and interact with other users online. The aforementioned practices, even if they are technically legal, can be harmful to people's mental health, the development of children and teenagers, and even to democracy. Conversely, practices that are prohibited by certain legislations[5] such as hacking, data breaching, forgery, fraud, child pornography, and terrorist activity. These types of practices, while harmful, fall outside the scope of what is considered online harmful practices and content in this essay, since they tend to be already regulated in a straight-forward manner.

---

3. White Papers are, according to the UK Parliament's Glossary "policy documents produced by the Government that set out their proposals for future legislation". The Online Harms White Paper consists of recommendations and suggestions by the UK Government to tackle online harm (Department for Digital, Culture, Media & Sport, 2020).

4. Paul Bernal,"Response to Online Harms White Paper", *Paul Bernal's Blog*, July 3, 2019, available at https://bit.ly/3VroZQk.

5. For instance, see the Budapest Convention on Cybercrime (ETS number 185), acceded by 68 States, which promotes measures against the following practices: i) illegal access, ii) illegal interception, iii) data interference, iv) system interference, v) misuse of devices, vi) computer-related forgery, vii) computer-related fraud, viii) offences related to child pornography, ix) offences related to infringements of copyright and related rights.

Online harm: The unavoidable problem to face and its challenges

The slow but steady recognition of online harm

During the last decades, there has been an evolution in the overall notion of what is or is not permissible online. This went from a very libertarian view[6] of what should be permitted online, to progressive regulation shown by some countries[7] and rising voices in the literature condemning the harm that the content provided by users on unmediated platforms and websites can bring to other users worldwide. As Citron (2019: 122) stated:

> A decade ago, any suggestion that the law should be brought to bear against cyber harassment was met with hostility. Commentators argued that people should be allowed to say anything they wanted online because their comments constituted speech. A common view was that the internet would cease to foster expression if law intervened.

Flashforward to today's latest landmark on the subject with the EU's Digital Services Act,[8] which, according to Vestager is linked to the idea that "what is illegal offline

6. John Perry Barlow, «Declaration of the Independence of Cyberspace», *Electronic Frontier Foundation*, February 8, 1996, available at https://bit.ly/4aiBm5o.

7. For instance, in the United States, Section 230 of the Communications Decency Act of 1996 regulates "protection for private blocking and screening of offensive material" and establishes in §230(b) the policies undertaken by the United States among which it is worth mentioning: §230(b)(4) "to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material"; and §230(b)(5) "to ensure vigorous enforcement of Federal criminal laws to deter and punish trafficking in obscenity, stalking, and harassment by means of computer". Furthermore, §230(b)(A) ensures that providers or users are not liable for "any action voluntarily taken in good faith to restrict access to or availability of material that the provider user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected". On the other hand, in the EU, article 3 of the Directive 2000/31/EC of the European Parliament and of the Council of June 8, 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce), allows Member States to restrict the freedom to provide information society services from another Member State if some conditions are fulfilled, one of them being the measure necessary for the "prevention, investigation, detection and prosecution of minors and the fight against any incitement to hatred on grounds of race, sex, religion or nationality, and violations of human dignity concerning individual persons".

8. Regulation (EU) 2022/2065 of the European Parliament and of the Council of October 19, 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). It was published in the Official Journal of the European Union on October 19, 2022, and came into force on November 16, 2022. It establishes rules for very large online platforms and search engines, online platforms, hosting services and intermediary services. The strictest rules apply to Very Large Online Platforms (VLOPs), that is, platforms that have more than 45 million monthly users in the EU. As of April 25, 2023, the following platforms were designated VLOPs by the Commission: i) Alibaba Aliexpress, ii) Amazon

is equally illegal online" and responds to the increase of online traffic, which renders necessary to create rules "that put order in the chaos".[9]

The fact that online harmful conduct is being increasingly recognised and some regulatory strategies have been adopted in order to tackle them, is a positive step towards a safer online environment. However, it remains relevant that the authorities that come up with said strategies acknowledge that new ways of online harm are likely to keep emerging as more and more people engage in the online world. In this respect, we highlight a statement present in the Online Harms White Paper regarding the list of harmful content that determines the scope of the document, a list that is thought to be "neither exhaustive nor fixed. A static list could prevent swift regulatory action to address new forms of online harm, new technologies, content and new online activities".[10]

Victim's profile: Who are the main targets of online harm?

Since there are different types of practices that can be conceptualized as online harm, it is helpful to make a distinction between the types of conduct to make an assessment regarding the profile of victims. There are certain types of online harm, such as misinformation, to which we are all arguably likely to be victims of, whether by directly engaging in social media, platforms, or forums, or by receiving said content from other users who share it (for instance, via direct communication using instant messaging applications such as WhatsApp or Telegram).

When it comes to harassment or sexualized abuse, according to Chemaly (2019), there are some groups that are particularly prone to being victims of them due to belonging to certain groups based on gender, race, religious or political beliefs. In this sense, she argues that

> women […] are the majority of targets of sustained and sexualized abuse, stalking and online harassment linked to both anonymous and intimate partner violence. While anonymity is a contributor, is not itself the sole cause. A large percentage of women know their aggressors (Chemaly, 2019: 150).

Store, iii) Apple AppStore, iv) Booking.com, v) Google Play, vi) Google Shopping, vii) Instagram, viii) LinkedIn, ix) TikTok, x) Twitter (now X), xi) Wikipedia, and xii) Youtube. Considering the impact the VLOPs can have on users around the world, amongst their duties they must tackle illegal content in their platforms and avoid risks regarding their users fundamental rights, such as discrimination and violence amongst users.

9. «Statement by Executive Vice-President Vestager on the Commission proposal on new rules for digital platforms», European Commission, December 15, 2020. Available at https://bit.ly/4aoOKLw.

10. See Department for Digital, Culture, Media & Sport (2020).

Franks (2019: 137), also argues that, regarding the victims of online abuse,

> online abuse jeopardizes victims' physical safety, employment opportunities, educational achievement, personal relationships, and psychological health. […] The internet all too often serves as a force multiplier for the harassment of women, lowering the costs of engaging in abuse by providing abusers with anonymity and social validation, while providing new ways to increase the range and impact of that abuse.

Accordingly, the UN in a study also showed that 73 % of female journalists declared having suffered from a type of online harm (Posetti and others, 2020). The explanatory memorandum in the proposal for the EU's Digital Services Act (DSA) mentions the issue of targeted conduct by claiming that "specific groups or persons may be vulnerable or disadvantaged in their use of online services because of their gender, race or ethnic origin, religion or belief, disability, age or sexual orientation" (European Commission, 2020: 12). Moreover, one of its aims is to serve as a mitigator of discriminatory risks, by overall protecting the right to human dignity online (European Commission, 2020).

The mentioned targeted groups of online victims are just an example of how the online experience and interaction among users are not the same for all who dare endeavour in cyberspace. While a deeper analysis of the reasons behind this targeting of certain groups online is outside the scope of this essay, we acknowledge it as an issue that creates problems for the measures taken so far, since they may not always recognise who are the main targets of online harm. In this sense, we believe the DSA moves in the right direction by directly acknowledging the problem.

The role of platforms and algorithms in online harm

When it comes to the role that platforms may have regarding the regulatory measures that can or should be taken to prevent and tackle online harm, it is important to mention at this point that online harm presents itself as a significant conflict of interest to such platforms, given that "harassing content is also a matter of corporate profitability. Abuse is emotionally resonant and it generates, in supporters or objectors, user activity. User activity means data, which is the lifeblood of online businesses" (Chemaly, 2019: 150).

In this sense, the business models of social media platforms, whether two-sided or multi-sided, that provide free services to users and rely on advertisers as their main source of income, might not be completely persuaded, at least from an economic standpoint, to effectively tackle online harmful conduct. This can be exacerbated by the use of algorithms, as some authors have argued:

> Inflammatory content draws more attention than uncontroversial topics, the general public has become desensitized to derogatory language. This is not surprising; however, algorithmic sensationalism amplifies derogatory messages in social media networks (Ascher and Noble, 2019: 170).

Moreover, this ill incentive for platforms may have as its cause the use of certain advertisements, as a whistle-blower and former Facebook employee described, in the sense that there is a structure of "subsidising hate" given that "hateful political ads were five to ten times cheaper for Facebook's customers, compared with empathetic or compassionate ads".[11]

Overall, it is safe to say that initiatives like the DSA with its substantial fines, aim to change the incentives faced by platforms in order to make the Internet and social media safer places, and it could serve as an important landmark that is later replicated by other countries and jurisdictions.

Online anonymity: A double-edged sword

Anonymity and its irruption in the online world

In the "offline world", it is difficult to envisage the same level of anonymity that appears to be so pivotal to the very existence of the Internet. When people incur a crime or harmful conduct, the concealment of their identity can precisely entail the annulment of their sense of accountability which is, in turn, an essential element of the very conduct. On the contrary, the unconcealment of their identity acts as a deterrent fuelled by the fear of socially or legally imposed sanctions. Quite the opposite occurs in the online realm, where anonymity is universally acknowledged, sought after, and even protected.

Lessig illustrates the effects of online interaction under real names, claiming that people are more likely to think before speaking and ensure being right before stating something as definitive. Furthermore, he mentions the influence of the "online community" on individual users, constraining them, and judging them with the consequence of being unable to escape the link to what one said. Moreover, he argues that "responsibility was a consequence of this architecture, but so was a certain inhibition" (Lessig, 2006: 95).

Anonymity as an online rule has been traced back by some authors,[12] originating from the cyberlibertarian view of the Internet by people such as John P. Barlow, who asserted in his *Declaration of the Independence of Cyberspace* that "legal concepts of property, expression, movement, and context do not apply to us. They are based on matter. There is no matter here".[13]

From this declaration, anonymity or pseudonymity —the absence of "matter" or "identity", in Barlow's words— can be easily regarded as essential to the Internet itself,

11. Haugen Frances, «Facebook whistleblower warns UK and EU to do more to control online harm», *The Financial Times*, November 9, 2021, available at https://bit.ly/3ILvB4c.

12. See Franks (2019: 137-149).

13. John Perry Barlow, «Declaration of the Independence of Cyberspace», *Electronic Frontier Foundation*, Febuary 8, 1996, available at https://bit.ly/4aiBm50.

to its virtual nature. This fact comes from the impossibility of physical persons to enter cyberspace in material form. This impossibility implicitly entails a dissociation between online and offline behaviour. Contrary to the belief on this dissociation, Gelber and Brison (2019: 12) have argued that:

> The internet is a place of material reality, not ethereal thought that is impervious to concrete consequences for those involved in using it, maintaining its platforms, or working in the industry surrounding it. This material reality stands in direct contradiction to Barlow's assertions in his Declaration.

The aforementioned dissociation between "material reality" and "ethereal thought" not only affects how people behave online but also the treatment and general perception of their behaviour, especially when it is harmful to others. The foregoing dissociation, present for several years, may have entailed a relative tolerance of online hate speech, sexual harassment, cyberbullying, doxing, and misinformation, among other harmful conducts, which have, according to some authors, been linked to the lack of accountability that comes attached to acting anonymously online.

This critique, of which anonymity online has been the subject, will be analysed below. However, we will previously refer to certain views in the literature that consider the privacy of users as one of the main reasons to defend anonymity. As we will see, and following Moore, these two features of anonymity —lack of accountability as a downside, and privacy for users as an upside— refer to different dimensions of anonymity. Hence, as they are of different dimensions, we will analyse them separately.

Is online anonymity even real? Brief analysis under the scope of data analytics and surveillance capitalism

As mentioned before, this essay follows Moore's three-dimensional approach to online anonymity. We will briefly refer to the dimension of traceability of anonymity, since privacy and the problem of traceability that users are increasingly being aware of, are perhaps one of the main defences for anonymity online. As Moore puts it,

> anonymity online, in the sense of "conducting one's affairs, communicating, or engaging in transactions…without one's name being known", is undermined by technologies that have made it possible to track or piece together the real identities of citizens online even when they are withholding their names or using pseudonyms (2018: 169).

This practice developed by companies such as Google or Facebook is located at the core of their business models, which is relevant since it has enabled these platforms to acquire their current positions in their respective markets as seemingly unassailable by rivals or authorities.

On the other hand, some of the heads of these platforms claim that this sort of digital market practice is intended to improve the users' experience in the platform itself. As Zuckerberg puts it,

> people consistently tell us that if they're going to see ads, they want them to be relevant. That means we need to understand their interests. So based on what pages people like, what they click on, and other signals, we create categories […]. Although advertising to specific groups existed well before the internet, online advertising allows much more precise targeting and therefore more-relevant ads.[14]

While this claim taken by itself does not sound unreasonable, once placed under the light of what surveillance capitalism really is it tends to seem like a "tip-of-the-iceberg" excuse when we consider, for instance, Zuboff's description of surveillance capitalism, which

> violates the inner sanctum, as machines and their algorithms decide the meaning of your sighs, blinks and utterances; the pattern of your breathing and the movement of your eyes; the clench of your jaw muscled; the hitch in your voice; and the exclamation points in a Facebook post once offered in innocence and hope (2019: 10).

Therefore, and following Moore, there appears to be a connection between the ability of governmental and private actors to trace users that can potentially constrain online communication due to the risk of exposure and retaliation for speech against powerful actors, which may be seen as a downside for traceability, but "while there are good reasons to resist traceability, there are also good reasons to want users to be traceable, such as identifying those who make threats or engage in hate speech and abuse" (Moore, 2018: 169).

While in this context the stakes may not seem that high —after all, being subject or predisposed by targeted ads does not necessarily override our freedom to decide to buy a product or not—, when these techniques are employed by powerful companies that have access to increasing levels of data, or governments take hold of these information and surveillance practices, anonymity seems to be a shield to face these entities' intervention, manipulation and surveillance of the daily lives of users. In this sense, Froomkin has aptly stated that:

> Anonymity may turn out to be the only tool available to ordinary people that can level the playing field against corporations and governments that might seek to use new data processing and data collection tools in ways that constrain the citizen's transactional or political freedom. […] Larger and faster database processing techniques combined with the ever-increasing quantity of personal data available about indivi-

14. Mark Zuckerberg, «Understanding Facebook's Business Model», *Meta*, January 24, 2019, available at https://bit.ly/3x4sO3F.

duals make it possible for both governments and private organizations to construct personal profiles based on transactions, demographics, and even reading habits of most citizens (1999: 113).

Relationship between anonymity and online harm

While anonymity, being ethically neutral (Reader, 2012) cannot be accused of being the only cause of online harm (Chemaly, 2019: 150), it has been addressed as an enabling factor for it since the origins of the Internet (Forestal and Phillips, 2019: 573). It even has been connected to pre-internet harmful practices like hate speech. This conduct, as Baker has argued, is a natural co-factor of anonymous communication regardless of where it takes place, given that anonymity "severs social and personal accountability" (2014: 166).

As a result, what could be considered as neutral or positive features and consequences of anonymity —namely disinhibition and privacy of users—, when placed in an online harm scenario, it can easily acquire a negative connotation by becoming a source of empowerment for harm, followed by a cloak of protection for the infringers or cyber-attackers who lack or have little accountability and can now engage in the ubiquitous space of the Internet, having access to an unprecedented level, immediacy and perpetuity of online content (Weinstein, 2019: 52). This idea relates also to what Lessig stated: "Just as anonymity might give you the strength to state an unpopular view, it can also shield you if you post an irresponsible, or slanderous, or hurtful view" (2006: 104).

Pseudonymity on the other hand has not been left out from scrutiny or criticism by real-name advocates. They have addressed the deceptive potential of pseudonymity as an enabling factor for the spread of fake news (Ascher and Noble, 2019: 170) and the fact that platforms such as X, formerly Twitter, can provide a "false sense of security that users have in their anonymity while accumulating social power under the guise of pseudonyms" (Ascher and Noble, 2019: 170).

We agree with Moore in the sense that "many contemporary critics of online anonymity share this framing of anonymity as a means to evade accountability for one's actions" (2018: 169). According to this author, this lack of accountability is what seemingly tends to hide people and enable them to engage in online harmful conduct such as "harassment, threats, bullying, defamation, lying, reputational damage, misogyny, and provision of false information, and protects from legal sanctions" (Moore, 2018: 169). He also mentions a more general effect regarding the degradation of public discourse and debate, which goes beyond the harm suffered by a person or group of persons but can actually degrade what is thought to be a marketplace of ideas and democracy.

However, it is important to mention that some authors, like Bernal, have argued against real-name policies or de-anonymizing tools since they might be counterpro-

ductive to deter *trolls*.[15] In this sense, he argues that "it may be that having a real name displayed emboldens trolls, adding credibility and kudos to their trolling activities".[16] This argument, in our opinion, does not provide a counterargument to the link between anonymity and lack of accountability as noted by other authors, and is seemingly encapsulated in an *ex-post* consideration of anonymity, that is, the general rule for a de-anonymizing tool to take place is that the users were anonymous in the first place.

Furthermore, the level of deterrence of a certain measure does not seem to be a sufficient argument to stop acting to avoid or prevent harm, as can be seen analogously in the scope of criminal law, which will still try to tackle crime even if there are undeterrable persons that do not respond to the threat of punishment. Finally, whereas Bernal's argument may be true for conducts such as *trolling*, there are other online harms that might be more systemic, such as disinformation, in which case, when real-name users incur in the spread of fake news for example, their credibility or discursive ability might be seriously compromised in front of their peers. Hence, the threat of losing credibility might be a proper deterrence to incur in such a practice and being more thorough before sharing unchecked information.

Relationship between anonymity, online harm and freedom of speech in the United States and freedom of expression in the European Union: A brief analysis

When it comes to linking freedom of speech to anonymity, as Froomkin states, the US First Amendment does not expressly guarantee the right to be anonymous as such, rather it

> guarantees of free speech and freedom of assembly (and whatever right to privacy exists in the Constitution) have, however, been understood for many years to provide protections for at least some, and possibly a great deal of, anonymous speech and secret association (1999: 113).

Accordingly, Kaminski argues that the goal of the First Amendment, apart from protecting speakers, is to provide "value to society as a whole, through the creation of a vibrant and egalitarian marketplace of ideas" (2012: 815).

15. Bernal characterizes two types of so-called internet *trolls*, after acknowledging that "it is not really clear what trolls actually are". The first kind of internet *troll* "puts out provocative information or says provocative things in places online where people are likely to respond"; and the second refers to "essentially anyone doing bad things on the internet, whether it be deliberate provocation in discussion or the worst kind of racism and misogyny and threats of violence, death or rape —or even political manipulation" (Bernal, 2018: 196).

16. Paul Bernal,"Response to Online Harms White Paper", Paul Bernal's Blog, July 3, 2019, available at https://bit.ly/3VroZQk.

As for the EU's view of this right, conceptualised as freedom of expression, it is contained in the European Convention of Human Rights (ECHR), article 10. It does not refer expressly to anonymity either, but it

> include[s] freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. […] The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and necessary in a democratic society, […] for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others.

If we link this conceptualisation of freedom of expression to Moore's dimensions of anonymity, we can see that moratorium on interference by public authority relates to the traceability dimension, and the reference to the duties and responsibilities that come from such a right includes the protection of reputation or rights of others. This last idea can be linked to the connectedness dimension of anonymity, which revolves around the concept of reputation, whether it is in a certain platform, or between different platforms and between the offline-online circles that a person engages with.

Thus, the undesirable consequences of online anonymity continue to jeopardize the values that were sought by anonymity in the first place, if considered as a corollary of freedom of speech, namely equality, democracy, civic inclusion, the privacy of users, and its associational function (Gelber and Brison, 2019: 12; Ascher and Noble, 2019: 170; Forestal and Phillips, 2019: 573). Anonymity in its traceability dimension can therefore allow for the engagement of marginalized users in public speech, promoting a richer marketplace of ideas and protecting new —but sometimes controversial, and other times unpopular— points of view (Forestal and Phillips, 2019: 573). And yet, according to PoKempner,

> as more and more of our lives move online, the rights most under threat are those that seemed at first to epitomize the internet: freedom of expression, access to information, and privacy —critical components of our self-awareness, autonomy, and dignity (2019: 224).

Regarding the complexity of online harm, it has been suggested that it has been allowed due to the idea that the solution and correction of harmful practices could prove to be worse than the initial problem, and the costs caused by regulation could never outweigh the damage to freedom of speech on the Internet (Haines and others, 2012: 765). Therefore, the "benefits of legal action were far outweighed by the costs to free expression. For the good of free speech, abusers needed to be let alone" (Citron, 2019: 122) or, as Weinstein argues: "As obnoxious, harmful, and discriminatory as cyber harassment may be, it nonetheless constitutes expression" (2019: 52). Franks (2019: 137)

has pointed out that there are types of abuse that were augmented or originated by the Internet and could be hard to address due to possible accusations of censorship.

In this sense, we believe that the key factor to consider, as with all human rights, is that they do not work as absolutes but come with certain limitations. Thus, Citron addresses this point by stating that:

> First Amendment protections and free speech values do not work as absolutes. There are speech interests beyond those of the harassers to consider. Rather than working against speech interests, intervention against online abuse would secure the necessary preconditions for free expression while safeguarding the equality of opportunity in the digital age (2019: 122).

Therefore, not everything that is spoken —whether orally or by written word— deserves to be the object of protection under this fundamental right, especially when they scarcely contribute to conversations on culture or politics, and above all when they entail harm to others (Citron, 2019: 122). In this sense, John Stuart Mill's view of debate and counterarguments could be achieved by freeing speech as much as possible, which is not really possible under the threat of online abuse suffered by targeted individuals who are therefore prevented from achieving their digital citizenship (Citron, 2019: 122). Or, as Citron puts it, "when a person's self-expression is designed for the purpose of extinguishing another person's speech, it should receive no protection" (2019: 122).

One counterargument to this view that can be found in literature, is that limiting free speech or policing the expression of certain views is considered by some to be tantamount to policing thoughts, given the "assumption that speech deserves special protection because it is more like thought than conduct" (Gelber and Brison, 2019: 12). Moreover, there are those who claim that words cannot injure as actions do. To this view, Gelber and Brison argue that "if it were really the case that words could never cause injury, a free speech principle would be otiose, adding nothing to a general principle of liberty grounded in the harm principle" (2019: 12). They furthermore state that speech is "a physical phenomenon, being instigated by agents, expressed by agents, and having physical effects on its listeners, effects that can be caused by the content of the speech" (Gelber and Brison, 2019: 12). As such, it should be prone to be regulated, however cumbersome this might prove in practice, whether in an offline or online dimension.

As a final thought on this subject, we believe that the concept of freedom of expression needs to be assessed under the reality of our time, adapting it to the ubiquity of the Internet, which has expanded the means for expression in a way that was probably unassailable at the time of the writing of the First Amendment and the ECHR. In this sense, Bezanson, argues that free speech is:

> Perhaps uniquely among the constitutional guarantees, a social guarantee that must always change faces as it reflects changing social conditions. It was born in a largely oral and infant-print environment. […] The culture was rural and communication over large distances was difficult, not instantaneous and cheap, as today. And so on. The speech guarantee therefore *must* change over time, just as the meaning of communication and the instruments it relies upon will change (2012: 237).

Is the regulation of anonymity the answer?

As seen throughout this essay, there are several dilemmas and weighing of benefits and costs before finding a solution to online harm by addressing changes in anonymity policies.

To answer the question of whether the regulation of anonymity is the answer, we will follow Lessig's modalities of regulation which constrain behaviour, namely: the law, social norms, markets,[17] and architecture (Lessig, 1999: 92). We will briefly examine each of these categories and assess whether anonymity does or should play a role in them.

Law modality

Regarding the law, Lessig mentions that its role in regulating cyberspace is by ordering people to behave in certain ways under the threat of punishment in case of disobedience (Lessig, 1999: 92). In that sense, what began as the law regulating against copyright, defamation, and obscenity (Lessig, 1999: 92), has been increasingly switching to attempt —not without controversy and not necessarily in an efficient manner— to prevent online harm.

An example that fits into this category could be the Digital Services Act and the adopted resolutions that

> include a strong call for […] protecting fundamental rights in the online environment, as well as online anonymity wherever technically possible. They call for transparency, information obligations and accountability for digital services providers and advocate for effective obligations to tackle illegal content online (European Commission, 2020: 1).

As can be seen, it addresses online anonymity only in general terms, acknowledging that it should be protected alongside fundamental rights, and therefore maintaining an equilibrium between the two notions, although the protection of anonymity seems to be more nuanced and subjected to where it is technically possible.

This regulation uses a broad concept of "illegal content" and places a strong focus on the duties of digital platforms, such as their roles in "content moderation", for which

17. Regarding the market modality, we refer to what has been stated about surveillance capitalism.

they must take measures to tackle illegal content or information that is incompatible with their terms and conditions. Among the measures to be taken, the DSA includes the "demotion, disabling of access to, or removal thereof, or the recipient's ability to provide that information, such as the termination or suspension of a recipient's account" (European Commission, 2020: 45). The adoption of these features by platforms will in turn interact with Lessig's code or architecture modality covered later in this essay, properly incentivised by the DSA's measures regarding non-compliance.

Another manifestation could be the United States Department of Justice's legislative reform proposal, in order to "modernize and clarify the immunity that 47 U.S.C. §230 provides to online platforms that host and moderate content" (Barr, 2020: 1) and to "take into account the vast technological changes that have occurred since the Communications Decency Act of 1996 was passed to incentivize online platforms to better address criminal content on their services and to be more transparent and accountable when removing lawful speech" (Barr, 2020: 1). This initiative, among other objectives, aims to incentivize online platforms to properly address illicit content by denying immunity to platforms from civil liability regarding users who suffer the consequences of particularly harmful practices.[18] Moreover, regarding content moderation by platforms, this reform would potentially promote open discourse and transparency, by changing the current system of immunity for platforms that remove content considered by them to be "objectionable", as per the concept used by Section 230(c)(2). This highly subjective term would be replaced by "unlawful" and that it "promotes terrorism", which would limit the platforms' ability "to remove content arbitrarily or in ways inconsistent with its terms or service by deeming it 'objectionable'".[19]

Social norms modality

According to Lessig (2006), social norms also regulate behaviour in cyberspace, and that has been the case since its origins. As is the case in jurisdictions with limited or inefficient regulation, what can happen in social platforms that show a slow level of responsiveness to online harm is the emergence of vindicative hacking, trolling, or hate speech as a way of taking justice into their own hands. As PoKempner said, "rights can be protected by the design of technology as well as through enabling individuals and communities to push back against undesirable speech" (2019: 224). Some problems with that statement may arise when said "pushing back" becomes retaliation via disproportionate or harmful means, when measures are taken without a proper assessment

18. Child exploitation and sexual abuse, terrorism and cyber-stalking are currently the proposed exemptions for immunity in case a platform fails to properly tackle content related to these practices.

19. Department of Justice's Review of Section 230 of the Communications Decency Act of 1996, available at https://bit.ly/3VvoHYm.

of the facts, or when taken against the wrong user.[20] In this sense, digital vigilantism or digilantism is an example of measures taken by users as a response to the breaking of online or offline social norms, which can be as harmful as the practice it is trying to regulate, either to deter or to vindicate. Practices such as doxing, hacking, or harassing, when adopted by users as a response to an initial rule broken or provocation, can leave the platform and its users in a constant state of self-censorship which also deteriorates free speech or the free exchange of ideas.

Overall, there is the possibility that the more effectively cyberspace is regulated by the modality of law, the less need there will be for *digilantism*, or other types of regulation enforcement by peers, especially when is carried out by harmful means. Furthermore, Véliz proposes an eclectic solution regarding the breaking of social norms, stating that they should not be too rigid in order to allow benefits such as evolving views and changes, for which "deviation of social norms should not be too costly" (2019: 633). She proposes pseudonymity as a tool to obtain an equilibrium between social norms and the censorship that could come as a consequence of them by proposing that:

> In instances in which we, as a society, are fairly certain that our norms are correct, such as the case of death threats, speech should be very costly (e.g. causing one to lose the privilege of anonymity, facing legal consequences); in cases that inhabit grey zones, such as the debate about the moral status of animals, speakers should be made to face criticism, but in a way that does not make it too costly to express a defendable view that may be widely accepted in the future, thanks to speakers like them (Véliz, 2019: 633).

Architecture (or code) modality

This modality is characterized by Lessig as an analogy of how architecture constrains us in real space, such as "railroad tracks that divide neighborhoods, bridges that block the access of buses" (1999: 92). In cyberspace, the "software and hardware that make cyberspace the way it is, constitutes a set of constraints on how one can behave" (Lessig, 1999: 92). He mentions the role of code writers on setting the features that constrain behaviours (such as eavesdropping) by allowing other behaviour (encryption).

This relates to the role of platforms in the developing of not only their terms and conditions, policies, and content moderation tools,[21] but also in the very structure

20. For instance, the "blue-check-marked" Twitter user Will Smith received multiple insults and hate messages during and after the 2022 Academy Awards. However, many mistook another user's account —a consolidated figure in the "gaming" world— with the Hollywood actor who, during the televised ceremony, displayed physical and verbal violence against the host.

21. For instance, warning users with a banner when they are about to retweet or share a news link that the platform recognizes from the absence of clicks that the user has not yet visited, as a tool to prevent the proliferation of fake news.

and features of their sites, the level of authentication requested from users, and the possibility of using anonymous accounts, among others. Some of these features will be intrinsically related to the main purpose for which the site or platform was created (for instance, connecting people with their friends in the case of Facebook, or connecting potential employers with employees in the case of LinkedIn), in which case anonymous or fake accounts, while possible, are against the purpose of the very goal of the platform. Other features will be adopted due to the influence of the law modality, which, if properly crafted, by placing incentives to which platforms respond, can entail cooperation between the two modalities of regulation, as opposed to competing between them (Lessig, 1999: 92).

However, regulation of content in general by platforms has appeared to be an erratic solution so far, apparently caused by the economic conflict of interest that it entails for them.[22] Furthermore, platforms' approach to content monitoring has not been exempt from problems such as bias,[23] or being abuser-friendly, and in most cases, the platform's response of taking down harmful content comes after the damage has already taken place. As mentioned, their level of positive response and adoption of the measures contained in legislation such as the DSA will probably depend on where the legislation puts the incentives for platforms, so that it is a viable option for them and their business models. Accordingly, the behaviour of the existing platforms towards straightforward legislation such as the DSA will determine its effectiveness as a regulatory measure or unmask its rapid obsolescence. Moreover, the behaviour of users and their compliance with social norms or lack thereof will also determine the effectiveness of both legislation and the platform's architecture in preventing online harm, whether it is by leaving them for other online spaces for fear of being authenticated, creating new accounts, or finding new ways of surpassing content monitoring.

Conclusion

Throughout this essay, we have explored the problem of online harm and anonymity as arguably being one of its enabling factors for which measures have been taken by users, platforms, and legislators, with different levels of involvement through the regulation of anonymity. The level of accuracy or effectiveness of these measures is interconnected,

22. See PoKempner (2019: 224).

23. It could be argued that a solution for bias would be initiatives such as Meta's Oversight Board, an extralegal independent body that either upholds or overturns content moderation decisions made by Facebook or Instagram regarding content shared by users. However, as Douek notes, there are further issues with such an enterprise, such as the lack of enforcement that this extralegal "court" and its rulings may have for platforms or users, or the creation of multiple parallel extralegal "courts", rendering the decisions of either body ultimately irrelevant. For instance, the so-called "Real Facebook Oversight Board" that was formed as a response to Meta's Oversight Board (available at https://bit.ly/43sGG3Y).

as Lessig's modalities of regulation interact, whether by cooperating or competing, and its full potential is yet to be discovered as online interaction continues to grow —with the correspondent increasing of social norms—, and different jurisdictions adopt legislative measures to address the problem of online harm whether by tangentially addressing them or by considering the accountability of users, or by focusing more on intermediary services and the platforms' monitoring role, as stated in the DSA.

Regarding online anonymity *per se*, its role in enabling online harm is a disputed area in literature and it is plausible to state that not all anonymous or pseudonymous users engage in online harmful activities, nor all real-name or authenticated users are innocent when it comes to harmful online practices. And yet, harmful conduct, in the absence of effective regulation, can lead to mental health issues, suicide, censorship, and the deterioration of democracy, among many other consequences. Conversely, anonymity, if it is not the enabling factor for disinhibiting perpetrators, can still serve as a cloak for the inflictors of such harms, and therefore aid and abet them. In turn, the lack of effective regulation can lead to disproportionate levels of retaliation via *digilantism*, the consequences of which can be worse than the initial harm that provoked it in the first place.

It also appears that while all the dimensions of anonymity display a bundle of costs and benefits, the calculation of which is utterly complex, it is by disaggregating anonymity in its different dimensions as Moore does, that it is possible to appreciate different levels of usefulness in anonymity at least in a theoretical standpoint. This exercise is far from futile since it can bring non-binary solutions, such as Véliz's (2019) proposal of pseudonymity as a compromise between real-name policies and anonymity of users, which provides a sufficient level of accountability that holds users liable and subject to consequences when incurring harmful practices, while also protecting free speech and the exchange of ideas in the marketplace of ideas.

It is also possible that all the problems caused by online harmful practices and their consequent responses by Lessig's (2006) regulation modalities, as well as the willingness or lack thereof from *netizens*[24] to adapt to them, will always remain linked to their own time and online experience. As of today, online users are a mix of digital natives and people who lived before the Internet was as omnipresent as it is today, and who can still remember it as a highly unregulated, Barlow-influenced territory. This can be reflected in a consequential resistance to new policies or regulations regarding cyberspace in general and the limitation of anonymity in particular.

As stated above, the effectiveness of the approaches given by the issue of online harm remains to be seen. But we believe that, as long as their goal and spirit are focused on erasing the dividing line between online and offline illegality by clearly defining what illegal conduct is, and by boosting accountability from the authors of online harmful conduct, the measures taken will be moving in the right direction.

24. Term used to describe a habitual or keen user of the Internet.

References

Ascher, Diana and Safiya Noble (2019). "Unmasking Hate on Twitter. Disrupting anonymity by tracking trolls". In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 170-188). Oxford: Oxford University Press.

Baker, Robert (2014). "Against Anonymity". *Bioethics 28*, 4: 166-169. DOI: 10.1111/bioe.12093.

Barr, William (2020). "Cover Letter to the Honorable Michael R. Pence, President of the United States Senate". Available at https://bit.ly/3VvoHYm.

Bernal, Paul (2018). *The Internet, Warts and All: Free Speech, Privacy and Truth*. Cambridge: Cambridge University Press.

Bezanson, Randall (2012). *Too Much Free Speech?* Champaign: University of Illinois Press.

Chemaly, Soraya (2019). "Demographics, Design, and Free Speech. How demographics have produced social media optimized for abuse and the silencing of marginalized voices". In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 150-169). Oxford: Oxford University Press.

Citron, Danielle Keats (2019). "Restricting Speech to Protect It". In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 122-136). Oxford: Oxford University Press.

Department for Digital, Culture, Media & Sport (2019). *Online Harms White Paper*. Available at https://bit.ly/49rlBIz.

European Commission (2020). Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. COM/2020/825 final. Available at https://bit.ly/3vQnTD3.

Forestal, Jennifer and Menaka Philips (2019). "The masked demos: Associational anonymity and democratic practice". *Contemporary Political Theory*, 19 (4): 573-595. DOI: 10.1057/s41296-019-00368-2.

Franks, Mary Anne (2019). "Not Where Bodies Live: The Abstraction of Internet Expression". In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 137-149). Oxford: Oxford University Press.

Froomkin, A. Michael (1999). "Legal Issues in Anonymity and Pseudonymity". *The Information Society*, 15 (2): 113-127.

Gelber, Katherine and Susan Brison (2019). "Digital Dualism and the 'Speech as Thought' Paradox". In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 12-33). Oxford: Oxford University Press.

Haines, Russell, Jill Hough, Lan Cao, and Douglas Haines (2012). "Anonymity in Computer-Mediated Communication: More Contrarian Ideas with Less Influence". *Group Decis Negot*, 23 (4): 765-786. DOI: 10.1007/s10726-012-9318-2.

Posetti, Julie, Nermine Aboulez, Kalina Bontcheva, Jackie Harrison, and Silvio Waisbord (2020). *Online violence Against Women Journalists: A Global Snapshot of Incidence and Impacts*. Paris: Unesco. Available at https://bit.ly/43nkhVq.

Kaminski, Margot (2012). "Real masks and real name policies: Applying anti-mask case law to anonymous online speech". *Fordham Intellectual Property, Media & Entertainment Law Journal*, 23 (3): 815-896.

Lessig, Lawrence (1999). "The Law of the Horse: What Cyberlaw Might Teach". In Madeleine Schachter (editor), *Law of Internet Speech* (pp. 92-100). Durham: Carolina Academic Press.

— (2006). *Code.* 2.ND ed. New York: Basic Books. Available at https://bit.ly/43nxj5w.

Moore, Alfred (2018). "Anonymity, Pseudonymity, and Deliberation: Why Not Everything Should Be Connected". *The Journal of Political Philosophy*, 26 (2): 169-192. DOI: 10.1111/jopp.12149.

PoKempner, Dinah (2019). "Regulating Online Speech. Keeping humans, and human rights, at the core." In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 224-245). Oxford: Oxford University Press.

Reader, Bill (2012). "Free Press vs. Free Speech? The Rhetoric of "Civility" in Regard to Anonymous Online Comments". *Journalism & Mass Communication Quarterly*, 89 (3): 495-513. DOI: 10.1177/1077699012447923.

Véliz, Carissa (2019). "Online Masquerade: Redesigning the Internet for Free Speech Through the Use of Pseudonyms". *Journal of Applied Philosophy*, 36 (4): 633-658. DOI: 10.1111/japp.12342.

Weinstein, James (2019). "Cyber Harassment and Free Speech: Drawing the Line Online". In Susan J. Brison and Katherine Gelber (editors), *Free Speech in the Digital Age* (pp. 52-73). Oxford: Oxford University Press.

Zuboff, Shoshana (2019). "Surveillance Capitalism and the Challenge of Collective Action". *New Labor Forum*, 28 (1): 10-29. DOI: 10.1177/1095796018819461.

About the author

María Francisca Ossa Monge is a lawyer who graduated from the Pontifical Catholic University of Chile. She also holds a Master of Laws (LLM) from The London School of Economics and Political Science and a Free Competition diploma from the Pontifical Catholic University of Chile. She is a professor of Economic Law at the Pontifical Catholic University of Chile. Her email is mfossa@uc.cl. https://orcid.org/0000-0002-5960-1519.

REVISTA CHILENA DE DERECHO Y TECNOLOGÍA

La *Revista de Chilena de Derecho y Tecnología* es una publicación académica semestral del Centro de Estudios en Derecho, Tecnología y Sociedad de la Facultad de Derecho de la Universidad de Chile, que tiene por objeto difundir en la comunidad jurídica los elementos necesarios para analizar y comprender los alcances y efectos que el desarrollo tecnológico y cultural han producido en la sociedad, especialmente su impacto en la ciencia jurídica.

DIRECTOR

Daniel Álvarez Valenzuela

(dalvarez@derecho.uchile.cl)

SITIO WEB

rchdt.uchile.cl

CORREO ELECTRÓNICO

rchdt@derecho.uchile.cl

La edición de textos, el diseño editorial
y la conversión a formatos electrónicos de este artículo
estuvieron a cargo de Tipográfica
(www.tipografica.io).