2020 | SEPTEMBER

SANTA CLARA PRINCIPLES

A CONTRIBUTION REGARDING TRANSPARENCY AROUND THE USE OF AUTOMATED TOOLS AND DECISION-MAKING

LAPIN

Submitted by:

Laboratory of Public Policy and Internet (*Laboratório de Políticas Públicas e Internet* - LAPIN)

Authors:

Felipe Rocha da Silva

Isabela Maria Rosal Santos

José Renato Laranjeira de Pereira (joserenato@lapin.org.br)

Mariane Andrade Moreira

Paulo Henrique Atta Sarmento

Thiago Guimarães Moraes

Cover Image:

Scott Webb, Unsplash

LAPIN
LABORATORY OF PUBLIC POLICY AND INTERNET

lapin.org.br
@lapin.br
/lapinbr
/lapinbr

About LAPIN

The **Laboratory of Public Policy and Internet (LAPIN)** is a pioneering think tank based in the Brazilian capital, Brasília. It brings together political scientists, lawyers, engineers and representatives from both public and private sectors to understand and support the development of public policies focused on the regulation of digital technologies.

Our mission is to investigate, analyse and understand the impacts of the Internet and new technologies on society and law, as well as to technically support public decision-makers in demands that involve digital themes, such as privacy, data protection, freedom of expression and respect for human rights on the Internet.

LAPIN has been advising members of the Brazilian Parliament by drafting technical notes and participating in public hearings. It has also assisted the Supreme Federal Court in constitutional actions regarding technology issues as *amicus curiae*. Moreover, LAPIN also raises awareness among society by organising events, such as workshops and webinars, to bring discussions on technology regulation to the wider public.

The think tank is currently at the forefront of the debates on the draft bill 2630/2020, which institutes the **Brazilian** Law of Freedom, Responsibility and Transparency on the Internet, the so-called "**fake news bill**", alongside other civil society organisations. LAPIN has helped highlight the theme to an international audience by moderating a debate about the theme at RightsCon 2020 under the panel "What is the Brazilian Fake News Bill", organised by Access Now and the Rights in the Network Coalition (Coalizão Direitos na Rede). The present submission benefits from the experience acquired throughout the debates of this draft bill to provide recommendations for the adoption of a principle related to algorithmic explainability. It is an honour to contribute.

Table of contents

I - Introduction

When surfing on the internet, our interactions leave traces in the format of data. Every click on a social network or search engine becomes a valuable piece of information which, collected in large amounts, allows for the creation of profiles of individuals which will be further applied by internet service providers (ISPs) to display personalised content for users.

As information processing increases, it has become easier to extract details about the life and personality of an individual from collected data which may at first sound inoffensive. From a simple search, it is possible to identify whether a woman is pregnant[1], a father is unemployed[2], or which political party one is likely to be affiliated with[3]. This type of personal data processing allows for the **personalisation of content** for each individual, which includes personalised targeted advertisements, content recommendation, and content ranking on social networks.

Some emblematic cases where the processing of personal data facilitated the spread of misinformation were the Cambridge Analytica scandal[4], the persecution of Rohingyans in Myanmar[5], and the Brazilian 2018 election campaign[6]. One thing in common in those cases was the dissemination of a large volume of false or decontextualized information in a personalized way to specific profiles on social networks to promote a particular political group or ideology.

[1] MAYER-SCHONBERGER, Viktor. Big Data : como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Tradução Paulo Polzonoff Junior. - 1. ed. - Rio de Janeiro : Elsevier, 2013, p. 40.
[2] Idem, p. 95
[3] HOWARD, Philip N. Lie Machines: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives. Yale University Press. New Haven, Londres. 2020.
[4] THE GUARDIAN. What is the Cambridge Analytica scandal? - video explainer. 19 mar 2018.
[5] HAHM, Jasmine. Facebook: Myanmar's Misinformation Megaphone. 23 dec 2018. Access: https://bpr.berkeley.edu/2018/12/23/facebook-myanmars-misinformation-megaphone/. Accessed 11 aug 2020.
[6] EVANGELISTA, R. & BRUNO, F. WhatsApp and political instability in Brazil. Internet Policy Review, Vol 1, Issue 4. 31 Dez 2019. Available at: https://policyreview.info/articles/analysis/whatsapp-and-political-instability-brazil-targeted-messagesand-political. Accessed on 11 aug 2020.

What we access online is largely the result of automated decision-making processes made by machine learning systems through inferences about our personality. Based on this data, systems are able to display customised content according to each individual, who has little or no autonomy over the profiling process. This form of use of personal data is a widely explored business model, capable of extracting predictions about who we are and what we might be thinking. Thus, transparency emerges as a fundamental concept for users to know which personal data are being processed and which inferences are made from those data.

Considering that social networks' content personalisation systems are responsible for most of what we access online, including disinformation, it is of utmost importance that the Santa Clara principles be expanded to include specific recommendations for transparency around automated tools and decision-making in the areas of content moderation, ad targeting and content recommendation. In this sense, **a fourth principle should be included to provide better guidance for companies about how their automated processing of data should be transparent.**

People should be able to easily recognize when their data is processed by automated tools. To achieve this, companies should, on the one hand, describe which data is collected, which inferences are made from this data, and the impact of personal data processing on what we consume. On the other, platforms should also enable users to control how content is personalised and moderated.

In the following sections we will provide details about how Santa Clara principles should be expanded in order to include provisions on specificities of automated tools and decision making for content personalisation. At the last session of this document, we propose the wording for a new Santa Clara Principle: the **Explanation Principle**.

II - Transparency in content personalisation

What we see online is largely the result of decisions made automatically by machines based on artificial intelligence from inferences about our personality based on how we act in the digital environment. Based on this data, it is possible to have a broad view about an individual's attitudes, preferences and behaviours[7], and thus customize content for each user, who has little or no autonomy over this profiling process.[8] Having some control on what content is offered to us directly involves knowing what data is being collected and which profiles are being made about ourselves.

Platforms such as social media, search engines and streaming services apply content personalisation systems to offer information that they consider most relevant to each user. The purpose of these techniques is often to anticipate or even influence behaviours, preferences and actions of individuals.

By transmitting content that concerns only the interests of a specific person, internet service providers end up restricting the diversity of ideas to which she has access, creating the so-called **echo chambers**, in which individuals have access only to content shared by users who think in a similar way, feeding back a system that always works on the basis of the same worldviews.[9]

As a result, **content personalisation often ends up creating new barriers to critical self-reflection** about one's own ideas, by preventing the individual from having access to information that contradicts her[10].

---

[7] HOWARD, Philip N. ***Lie Machines****: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives.* Yale University Press. New Haven, London. 2020.

[8] BOZDAG, Engin. ***Bias in algorithmic filtering and personalization***. Ethics and Information Technology, 15(3), p. 211. 23 Jun 2013. Available in: < http://doi.org/10.1007/s10676-013-9321-6>. Accessed on 11 August 2020.

[9] MITTELSTADT, Brent. ***Auditing for Transparency in Content Personalization Systems***. International Journal of Communication 10(2016), 4991–5002. Available in: <https://www.ijoc.org/index.php/ijoc/article/view/6267>. Accessed on 11 august 2020.

[10] MITTELSTADT, Op Cit., p. 4994.

For this reason, when exploited for political purposes, personalisation may impact the foundations of a democratic system, since they limit the user's access to additional information that may question already consolidated beliefs through filtering tools. The result of this process has been the **radicalisation** of populations in more or less consolidated democracies around the world.

This also opens the door for the dissemination of misinformative content. During the 2016 elections in the USA and the campaign for Brexit, individuals were segregated into different groups of supporters of a particular candidate or ideology according to criteria such as their political vision, their health history, their social class or their geographical location based on personal data available on Facebook. At that occasion, algorithms reinforced clickbait news, a new way to disseminate political misinformation[11]. Inside these "bubbles", messages with fake content were disseminated and shared much more frequently than other types of content.[12]

Several of these messages contained fake, sensational, extremist, conspiracy, deeply biased content or comment masked as news, elements which Oxford University professor Philip Howard calls "junk news".[13] The most critical issue is that participants in these groups often did not know that this content had been specially tailored to them based on the processing of their personal data.

This phenomenon happens in many different platforms, from content-sharing social networks to search engines and streaming services. However, these platforms usually are not transparent about the profiles they use to frame their users, and their opacity is also often reflected in the lack of relevant information about the automated decision-making process used in both the profiling and the content recommendation.

It should be noted that content individualisation generates greater engagement in social networks, which enables companies to profit from the attention of its users by offering ads in their platforms. Since targeted marketing is the major source of income

---

[11] BENKLER, Yochai; FARIS, Robert; ROBERTS, Hal. *Network propaganda: manipulation, disinformation, and radicalization in American politics*. New York, NY. Oxford University Press, 2018. Pp. 10-11.

[12] RESENDE, G. *et al.* ***(Mis)information Dissemination in WhatsApp: Gathering, Analysis and Countermeasures***. WWW '19: The World Wide Web Conference. Available in: <https://dl.acm.org/doi/10.1145/3308558.3313688>. Accessed on 11 august 2020.

[13] HOWARD, 2020, Op Cit.

for several companies that carry out profiling processes, having content with high rates of involvement of its users helps the companies sell these ads[14]. Nevertheless, these ads are not frequently flagged as marketing content, and users barely know that the content they are receiving has been selected for their specific profile based on personal data processing.

For this reason, internet service providers should allow for more transparency and explainability about their automated decision content recommendation systems. This would guarantee that users, researchers and regulators have a better sense of how content personalisation algorithms influence the delivery of information.

Companies should ensure that **active transparency mechanisms** exist, whereby the data subject has the right to actively request which data is used and how the automated decision process is done. But it is also essential to encourage **passive transparency** instruments, by which the data controller informs the user by default of the existence of automated decisions and how they occur. After all, if the person is not passively informed that automated decisions are taking place, how will he or she take the initiative to request any revision of them?

The General Data Protection Regulation - GDPR - provides provisions on this matter. In its articles 13(2)(f) and 14(2)(g)[15], the regulation determines that the following information must be provided spontaneously by the data controller, regardless of request, to ensure fair and transparent processing:

1. The existence of automated decision-making, including profiling;
2. Relevant information about the logic of the algorithms responsible for automated decision-making;
3. Impact and envisaged consequences of such decision making.

[14] ZUBOFF, Shoshana. **The age of surveillance capitalism: the fight for a human future at the new frontier of power.** 1st ed. New York: PublicAffairs, 2018. Pp. 93-97.

[15] Article 13(2)(f), GDPR - *the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.*
Article 14(2)(g), GDPR - *the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.*

Other crucial aspects for a better understanding of the logic of a system and a specific decision relate to the following questions[16]:

1. What are the main inputs taken into account in an automated decision? This refers to data such as the location of a subject, his physical and psychological characteristics, such as gender, race, sexual orientation, political affiliations and religious beliefs, or yet the fact that he or she has been profiled as part of a particular group or has recently interacted with a specific content;
2. What was the weight of each of these factors during decision making?;
3. Can any of these elements receive different weights in similar automated decisions?

These questions are essential to acknowledge what factors were determinant in a given automated decision, and consequently identify the biases of a system, both to enable understanding why certain information was shown to a user and to perceive what information is not being displayed to her.

It is worth noting, however, that not every decision should be explained, under penalty of rendering the service provided unfeasible. The decisions that must be explained are the ones that:

1. Have a significant impact on the rights of an individual other than the decision-maker;
2. Serve to prove the liability of the platform for a certain harm caused;
3. Have well-founded evidence of errors[17].

[16] DOSHI-VELEZ, Finale et al. ***Accountability of AI Under the Law****: The Role of Explanation*. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper, 2017, p. 3. Available in: <nrs.harvard.edu/urn-3:HUL.InstRepos:34372584>. Accessed on 11 august 2020.
[17] *Idem*, p. 4.

The information listed here must be presented to the user individually, since it involves the processing of personal data that must be protected, in plain language, but also in transparency reports.

Explainability of a specific decision should be presented inasmuch as it is necessary to understand how personal information affects what one accesses and how to guarantee the autonomy of the data subject over what information about him is being processed[18].

In addition, it is of utmost importance to provide mechanisms for empowering the data subject[19] by allowing her to review the automated decisions made about her personality. This is the main reason why data protection regulations should provide for mechanisms to challenge automated decisions, as provided e.g. by Article 22(3) of GDPR[20] or Article 20 of the Brazilian General Data Protection Legislation[21]. Moreover, human review and explainability should be encouraged in decisions that affect the rights of an individual.[22]

Some examples for how a platform may provide for more transparency about the logic of an automated decision-making system or may allow for the exercise of the individual's right to contest a decision are through opt-out instruments for profiling or accessing certain kinds of ads; transparency reports on profiling algorithms; explicit labels on targeted advertisement, which includes identifying the ones paying for it;

---

[18] *Idem*, p. 9.

[19] FERRETTI, Federico. **Data Protection and the Legitimate Interest of Data Controllers: Much Ado About Nothing or the Winter of Rights?** Common Market Law Review 51. United Kingdom. 2014. Pp. 850-81.
RODOTÀ, Stefano. **A vida na sociedade da vigilância - a privacidade hoje**. Rio de Janeiro: Renovar, 2008. Pp. 46 e 47

[20] Article 22(3), GDPR - *In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.*

[21] Article 20, LGPD - *The data subject has the right to request review, by a natural person, of decisions taken solely on the bases of automated processing of personal data that affects her/his interests, including decisions intended to define her/his personal, professional, consumer or credit profile or aspects of her/his personality.*
English version available at https://iapp.org/resources/article/brazils-general-data-protection-law-english-translation/. Accessed on 27 Aug 2020.

[22] GILLESPIE, Tarleton. **Custodians of the internet**: platforms, content moderation, and the hidden decisions that shape social media. New Haven, 2018. Yale University Press. Pp. 114-117.

information about the profiles in which the individual was labeled[23]; the right of rectification of inaccurate personal data[24]. After all, participation of the data subject is essential in the fight against disinformation. It is more likely that a well-informed person will be able to verify abusive action by the company. That is why it is necessary to ensure these transparency and action tools.

For these reasons it is recommended that a new principle to ensure transparency in the profiling process for the purposes of content ranking, ad targeting and any other form of content personalisation is adopted under the Santa Clara Principles in order to encourage platforms to adopt mechanisms that empower the data subject to avoid accessing misinformative content. The recommendations which we present at the end of this submission shall guide platforms to adopt measures compatible with its business model.

[23] BIONI, Bruno Ricardo. **Proteção de dados pessoais: a função e os limites do consentimento**. 2nd edition. Rio de Janeiro: Forense, 2019, p. 256.
Also in BENKLER, FARIS,ROBERTS, 2018. *Op cit*. P. 372. It mentions that the companies should also take responsibility to afford such instruments. It examples that Google [announced that it would publish a transparency report about who is buying election-related ads on Google platforms and how much money is being spent, a publicly accessible database of election ads purchased on AdWords and YouTube, with information about who bought each ad, and will implement in-ad disclosures—Google will identify the names of advertisers running election-related campaigns on Google Search, YouTube, and the Google Display Network via Google's "Why This Ad" icon](...).

[24] Article 16, GDPR - *The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement.*

III - Transparency in content moderation

Social networks face a huge challenge to ensure that users follow the ethical principles set out in their terms of use. How to make sure that an efficient control of published content is carried out when thousands of messages are published every minute?

In a report produced by *Cambridge Consultants*[25] under the request of the British communications regulator, OfCom, it was shown that the **monitoring of content transmitted on platforms would be impossible without the aid of artificial intelligence technologies** in both *ex-ante* moderation activities (i.e. before the content is published on the network) and *ex-post* (i.e. after the content is published on the network). Algorithms have been key in performing content moderation to identify improper content and ensuring a healthier online environment.

Automated (or semi-automated) **pre-moderation** is used for the **removal of illegal content** or content that conflicts directly with the platform's policies and is commonly used to detect "indubitable" illegal content, such as child abuse material or explicit violence. This activity is crucial to prevent inappropriate content from going online even before it is published by a user since the mere act of publicising a violent image or text is capable of causing harm to individuals. This is one of the reasons why companies like Facebook invest in improving algorithms that allow for the removal of inappropriate content before its publication even without human intervention.[26]

Post-moderation, on the other hand, is usually semi-automated, which means it requires the intervention of a human team to confirm the removal of inappropriate content.[27] Generally, content removed *ex-post* is the one that **depends on the analysis of contextual elements**, such as political views, cultural beliefs, historical events and

[25] "Use of AI in Online Content Moderation." *Ofcom*, Cambridge Consultants, 18 July 2019, www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-content-moderation.
[26] "F8 2018: Using Technology to Remove the Bad Stuff Before It's Even Reported." FACEBOOK. 2 May 2018, https://about.fb.com/news/2018/05/removing-content-using-ai/.
[27] CAMBRIDGE CONSULTANTS. Op cit.

local laws, and can hardly be removed automatically without posing a risk to freedom of expression.

For example, in 2016, Facebook suffered harsh criticism for removing a famous photo of a naked girl screaming and running after a napalm attack hit Vietnam in 1972.[28] The company later stepped back and republished the image. Although risks of false positives exist, artificial intelligence technologies can be used to assist human operators by signalling publications that possibly contain inappropriate content, so that they can decide on the removal of an image or text.

Although automation technologies are crucial for content moderation, one of its greatest issues is that, **in many cases, there is no accurate information about the period, amount and type of content that has been removed through automated decision making**. This information is valuable especially for performing an analysis of how misinformation is identified online. Moreover, the information provided to the authors of the removed content is also often vague and prevents a more detailed understanding of the reason behind the platform's moderation process.

An example of a transparency practice relates to the reports of compliance of YouTube's community guidelines, that informs the number of videos removed in a given period and highlights the cases of automated removals. The most recent report,[29] covering the period from April to June 2020, reveals that more than 11.4 million videos have been removed, of which 10.84 million (95%) were due to automated detection. However, the report does not distinguish between automated and non-automated removals for the reason of removal. If done so, this would allow users to identify which types of content are most commonly removed automatically, such as child abusive content, nudity, explicit violence and so on.

Therefore, it is essential that the Santa Clara Principles include a guideline for social media platforms to recognize the importance of **transparency in automated decisions**, whether at the time of notification of the user or in the preparation of reports. Rules should be in place regarding the need for this information to be

[28] Scott, Mark, and Mike Isaac. "Facebook Restores Iconic Vietnam War Photo It Censored for Nudity." The New York Times, 9 Sept. 2016.
[29] https://transparencyreport.google.com/youtube-policy/removals

disclosed in transparency reports and users to be notified when content removal occurs in an automated manner.

As previously stressed in Section II, any guidelines on the subject must be principle-based and technologically neutral, allowing each social network provider to adapt the **Explanation principle** according to the specificities of its business model. It is also important that **transparency reports include information about how and in what proportions automated decisions acted in the content moderation processes of the platforms**.

The **Explanation** principle, as a standard to promote algorithmic transparency, may equally fill gaps in other Santa Clara Principles. Its inclusion will thus strengthen systemically the whole structure, bringing higher practicality to it.

With regards to the **Numbers** principle, for instance, the application of algorithmic transparency to the disclosure of aggregated data related to content takedown and flagging in platforms' reports will most certainly turn these companies more accountable.

Regarding the **Appeal** principle, algorithmic transparency equally deepens users' capacity to comprehend the proper dimensions of how one's data is processed and, therefore, strengthens one's capacity to ponder whether or not it's necessary to contest specific decisions related to content flagging or takedown.

Lastly, algorithmic transparency prevents opacities where companies should inform its users about what content infringes its usage policies, evidently affronting the **Reports** principle. Higher transparency about the elucidation of how the users' posts infringe a platform's terms of use and more information of which technique was used to spot such disconformity will allow for a better implementation of this principle.

System's opacities not only generate information asymmetry between users and companies but equally disbalance the playing field in the competitive spectrum. "Technobabble", which is technical language that is difficult for ordinary people to

understand, often hide anti-competitive, discriminatory, or simply careless conduct[30], and only transparency is able to empower the users against such negative externalities.

Therefore, it is highly suggested that the Santa Clara Principles reflect concerns about transparency in content moderation. Effective moderation mechanisms can only be fully implemented with a transparent architecture, both related to the logic pursued by the algorithms applied in this process and to the interactions that concern this moderation.

[30] Pasquale, Frank. *Black Box Society: the Secret Algorithms That Control Money and Information*. Harvard University Press, 2016.

IV - Recommendations

We have described the importance, as well as possible paths, for implementing transparency mechanisms for both explaining automated decisions and providing further information about content moderation techniques applied by Internet Service Providers. We thus propose a fourth Santa Clara Principle on the terms described below: the **Explanation Principle.**

THE EXPLANATION PRINCIPLE

In order to ensure that individuals are on the control of the use of their personal data when accessing content online, as well as to identify how content is personalised and displayed by algorithms to them, companies should have mechanisms to guarantee the transparency of their automated decision making, which should include:

- Active transparency tools, allowing users to request access to information on automated decision-making and profiling made on them;
- Passive transparency tools, providing information by default on the logic of automated decision-making systems
- Disclosure of the profiling categories that have been assigned to the data subject;
- Identification (flagging) of advertisements, especially those that had been targeted to the user;
- Development of tools that allow for greater empowerment of the data subject, such as privacy dashboards and opt-in/opt-out buttons related to profiling and ad targeting.

Furthermore, companies should put in practice transparency mechanisms for how automated decision-making is used in content moderation, which should include:

- Greater disclosure of statistics about what content is removed by automated means and the motivation for the decision;
- More information about how automated decisions are used in "pre-" and "post-" moderation;
- Clear information about types of content that may be removed from the platform;
- Explanation of the content moderation process, and how human and non-human actors are involved in it;
- Reports including information about how and in what proportions automated decisions acted in the content moderation processes of the platforms; and
- Further information about the logics of the automated decision-making system applied, both for content personalisation and content moderation on transparency reports.