**Universidade de Brasília – UnB**
**Faculdade de Direito**

JOSÉ RENATO LARANJEIRA DE PEREIRA

# OPENNESS DOESN'T HURT: ENFORCING QUALIFIED MACHINE-LEARNING TRANSPARENCY FOR DATA PROTECTION THROUGH RESPONSIVE REGULATION

*Abertura não faz mal: promovendo transparência qualificada de sistemas de aprendizagem de máquina para proteção de dados por meio da regulação responsiva*

Brasília
2022

UNIVERSITY OF BRASÍLIA
FACULTY OF LAW

# OPENNESS DOESN'T HURT: ENFORCING QUALIFIED MACHINE-LEARNING TRANSPARENCY FOR DATA PROTECTION THROUGH RESPONSIVE REGULATION

Author: José Renato Laranjeira de Pereira

Supervisor: Prof. Dr. Márcio Iorio Aranha

Dissertation presented as a partial requirement for obtaining the title of Master under the Graduate Programme in Law at the University of Brasília, research line *Transformations in Social and Economic Order and Regulation*.

Brasília, 5 August 2022

**José Renato Laranjeira de Pereira**

**Openness doesn't hurt: Enforcing qualified machine learning transparency for data protection through responsive regulation**

Dissertation presented as a partial requirement for obtaining the title of Master under the Graduate Programme in Law at the University of Brasília, research line *Transformations in Social and Economic Order and Regulation*.

Presentation: 29 August 2022.

EXAMINATION COMMITTEE

_____

Prof. Dr. Márcio Iorio Aranha
(Supervisor – President)

_____

Prof. Dr. Michael Veale
(Member)

_____

Prof. Dr. Miriam Wimmer
(Member)

_____

Prof. Dr. Alexandre Kehrig Veronese Aguiar
(Substitute Member)

# ACKNOWLEDGMENTS

*Às minhas avós por suas histórias no sofá*

*Ao meu avô baiano por suas histórias sobre o sertão*

*Ao meu avô alagoano pelas histórias que não me contou*

*A Dilai por me ensinar a viver*

*Ao Atlântico*

*A Sophia,*

*meu Diadorim.*

[To my grandmothers for their stories on a sofa

To my grandfather from Bahia for his stories about the Sertão

To my grandfather from Alagoas for the stories he didn't tell me

To Dilai for teaching me how to live

To the Atlantic Ocean

To Sophia,

my Diadorim.]

**FICHA CATALOGRÁFICA**

Laranjeira de Pereira, José Renato
Lo Openness doesn't hurt: Enforcing qualified machine learning transparency for data protection through responsive regulation / José Renato Laranjeira de Pereira; orientador Márcio Iorio Aranha. -- Brasília, 2022.
183 p.

Tese (Doutorado - Mestrado em Direito) -- Universidade de Brasília, 2022.

1. machine learning. 2. transparency. 3. accountability. 4. responsive regulation. 5. artificial intelligence. I. Aranha, Márcio Iorio, orient. II. Título.

**REFERÊNCIA BIBLIOGRÁFICA**

LARANJEIRA DE PEREIRA, José Renato. 2022. **Openness doesn't hurt**: Enforcing qualified machine learning transparency for data protection through responsive regulation. Dissertação de Mestrado, Faculdade de Direito, Universidade de Brasília, Brasília, DF, p. 183.

*"'Seguro que fue un sueño', insistían los oficiales. 'En Macondo no ha pasado nada, ni está pasando ni pasará nunca. Este es un pueblo feliz.' Así consumaron el exterminio de los jefes sindicales."*

- Gabriel García Marquez, Cien años de soledad.

*"Eu sei que não morrer, nem sempre é viver. Deve haver outros caminhos, saídas mais amenas".*

- Conceição Evaristo, Olhos d'Água.

*"É necessário preservar o avesso,* você me disse. *Preservar aquilo que ninguém vê".*

- Jeferson Tenório, O Avesso da Pele

*(...) É assim que eu te escrevo*
*Nesse interlúdio*
*Em que te busco transparecer porque o mundo é opaco demais para mim.*

*E a minha travessia*
*Parecia demais*

*Purgatório*.

- Metamônico

# TABLE OF CONTENTS

# Abstract

Machine-learning (ML) models have been increasingly applied to make decisions that affect key aspects of people's lives. However, users and regulators are barely aware of how these models work, as only scarce information is disclosed by developers and operators on this matter. ML transparency emerges thus as a recurrent demand made by stakeholders for users to gain control over how much their lives should rely on judgements carried out by machines, for regulators to render those responsible for them accountable for incurred damages and for scholars to understand algorithms' impacts in society. This dissertation thus traces a comparative analysis on how the Brazilian and European data protection legal frameworks address ML transparency and assesses the adequateness of the responsive regulation theory's participatory strategies and incentives framework for promoting more intelligible systems.



Keywords: machine learning, transparency, accountability, responsive regulation, artificial intelligence.

# Resumo

Sistemas de aprendizagem de máquina (*machine learning*, ML) têm sido cada vez mais utilizados em processos de tomada de decisões que afetam aspectos-chave das vidas de pessoas. Entretanto, usuários e reguladores pouco sabem sobre como esses modelos funcionam, já que apenas informações escassas são divulgadas por seus desenvolvedores e operadores. A transparência dessas tecnologias surge assim como uma exigência feita por diferentes grupos de especialistas para que os usuários tenham controle sobre o quanto suas vidas devem depender dos julgamentos realizados por sistemas de *machine learning*, mas também para que reguladores responsabilizem os responsáveis por eles pelos danos que vierem a incorrer. Esta dissertação traça assim uma análise comparativa sobre como as leis brasileira e europeia de proteção de dados abordam a transparência de *machine learning* e avalia a adequação das estratégias participativas da teoria da regulação responsiva e de sua estrutura de incentivos para promover sistemas mais inteligíveis.



Palavras-chave: machine learning, transparência, responsabilidade, regulação responsiva, inteligência artificial.

**Figures**

**Tables**

**Abbreviations and Acronyms**

| | |
|---|---|
| AI | Artificial Intelligence |
| ANPD | *Autoridade Nacional de Proteção de Dados* (National Data Protection Authority |
| ANS | *Agência Nacional de Saúde Complementar* (National Agency for Supplementary Health) |
| CDR | *Coalizão Direitos na Rede* (Rights in the Network Coalition) |
| CNPD | *Conselho Nacional de Proteção de Dados* (National Data Protection Council) |
| DPIA | *Relatório de Impacto à Proteção de Dados Pessoais* / Data Protection Impact Assessment |
| GDPR | General Data Protection Regulation |
| GRID | Good Regulatory Intervention Design |
| LAPIN | *Laboratório de Políticas Públicas e Internet* (Laboratory of Public Policy and Internet |
| ML | Machine Learning |

## 1. INTRODUCTION

Artificial intelligence (AI) systems have occupied a crucial role in the functioning of most digital systems developed and adopted by both individuals and organisations. A fast-evolving family of technologies and a field of study, AI has been deployed, silently or not, in smartphones, autonomous cars, smart TVs, smart toys, social media platforms, surveillance cameras, and almost anything digital that reaches our hands in the 2020s.

AI algorithms have also been increasingly applied for supporting and making decisions that affect critical aspects of people's lives. They are being used to help businesses' and governments' decision-making processes in areas ranging from creditworthiness analysis and eligibility for welfare benefits to curating what political news we access on social media (CENTER FOR AI AND DIGITAL POLICY, 2020).

Two phenomena are among the most influential features for the growth in AI adoption in the past decades. The first relates to the ever-increasing amount of electronic data being produced every day. The second consists of the enhancement of data processing capacity in computational systems (KANDPAL, KRISHNAN & SAMAVEDHAM, 2012).

These trends had multiple consequences. Among them, one can mention the democratisation of access to new technology. As investments in companies such as Intel were driven mainly towards the increase in computer power, the processing capacity of computers increased exponentially. At the same time, the "price of new CPUs remained (fairly) stable; and cost of older technology dropped at (roughly) the same rate as the power of new processors rose" (KUNIAVSKY, 2010, p. 6).

The increase in the production and availability of data and the greater capacity to process more information were fundamental for the acceleration of AI development. As we have seen, data, and good-quality data, are essential to boost AI algorithms' performance, and being able to process it as fast and as efficiently as possible is a way to develop more efficient systems (GRÖGER, 2021).

Extracting value from data has been perhaps one of the most important competitive advantages of the companies that reshaped capitalism in the last twenty years, especially technology giants Google, Apple, Facebook and Amazon. Through their ability to commercialise and create new products from the tremendous amount of information - including personal information - they collect, they became the most valuable companies in the world. Most of this was also achieved by their capacity to develop advanced artificial intelligence systems that were proficient in extracting meaning from such data (VEALE, 2019). Important to note, nevertheless, that these players were also the main creators of what Shoshana Zuboff (2018) calls *surveillance capitalism*: a new economic order founded mainly on the hidden collection, accumulation and profiting from personal data through an architecture built towards modifying the behavior of individuals.

Due to the role AI has been playing in contemporaneity through the examples we have seen above, it has been gaining wider attention from society, especially as more information is disseminated about its influence on individuals' and groups' fortunes and how many systems have been replacing our agency in different spheres of our lives through automated decision-making processes. Before the growth of social media, for instance, news outlets were key curators of information in society. Now, some consider that AI systems governed by companies such as Facebook, Twitter and Google have taken the role of gatekeeping information as they have been the ones responsible for selecting content that we access online on their platforms (BUCHER, 2018). Another example is the application of AI systems for predictive policing, as a way to support police forces and policymakers in identifying the neighbourhoods in which crimes are more likely to be committed (DELOITTE, 2021). They are also the basis of many credit analysis systems and play a key role in determining our credit-worthiness (HEAVEN, 2021).

Nevertheless, as AI has been helping reshape how we access information, combat criminality and assess credit-worthiness, its pitfalls have also started to spread. Information leaked on the so-called *Facebook Files* showed how the company's algorithms may have contributed to the development of eating disorders in teenagers - and that the company was very well aware of that (ALTER, 2021). In other applications, dis-

criminatory bias has been found in face recognition systems, which were shown to fail more frequently when trying to identify people who were not white males (NIST, 2019) - and have even influenced police officers to mistakenly take into custody innocent people in Brazil (WERNECK, 2019; REIS, ALMEIDA, DA SILVA, DOURADO, 2021). Credit scoring systems based on AI have also been less accurate when assessing minorities' credit histories in the United States of America (HEAVEN, 2021).

Such hazards of AI systems indicate how they are far from neutral. Developed by human beings, they reflect our beliefs, prejudices, personal and societal backgrounds, as well as our hopes, fears, ambitions and biases, including a reproduction of systemic racism (SILVA, 2022; ALMEIDA, 2019). Hence, there is no other way to look at their promises and threats but as human creations, full of potential flaws, and not as all-encompassing solutions for the world's miseries.

And these pitfalls should be taken seriously. By selecting and ranking information to be displayed on social media, as well as by supporting decisions such as helping police officers decide whom to approach in the streets and banks to classify who is trustworthy enough to receive credit, these systems help to make the world appear in certain ways rather than others. As they influence how humans see the world, they also impact how we make our decisions.

### 1.1. The narratives surrounding AI

Narratives concerning artificial intelligence have evolved in directions as manifold as the technologies to which they refer. Some of them reflect a rhetoric of *inevitabilism* (ZUBOFF, 2019) that grew to become almost an ideology in the technological field. This narrative outlines that technological development as it happens to be is inevitable and indispensable for solving humanity's problems. This approach usually presents digital solutions as a "quick and flawless way to solve real world problems" (BBW, 2021), a mentality that Evgeny Morozov calls "techno solutionism" (MOROZOV, 2013), whereby individuals and organisations rely uncritically on technology to solve issues, without much consideration to its risks.

At the same time, there are multiple examples of reactions against the exercise of a sort of technological reign over society, calling attention to how any uncritical adoption of AI may have severe consequences on the rights and liberties of individuals and groups. Initiatives such as the Italian #NoStreamDay, which advocates against the opacity of AI-based content moderation systems in streaming platforms, as well as lawsuits and legislations in Brazil and Spain demanding gig economy companies to render their algorithms more transparent, represent some of the voices that call for further regulation of artificial intelligence systems. Organisations such as Big Brother Watch and Access Now, for instance, have been supporting the ban of facial recognition, an increasingly applied AI application. In another case, the European Commission has been driving efforts toward establishing ethical principles for artificial intelligence and even for the approval of a risk-based regulation through its AI Act proposal.

This cautious approach, referenced by Marda (2018, p. 1) as a "sobering narrative", emerged from the identification that these systems have a tendency to "not only imbibe, but also exacerbate existing human biases", and cannot escape from the fact that they are technologies susceptible of failures.

These divergent narratives reflect the influential role AI systems play in society. That's why governments, companies, academia and civil society have turned their attention both to AI's potential to enhance humanity's problem-solving capacity and the risks they impose on our welfare. Their risks drive society to conceive regulations for these tools that provide for effective, fair forms of accountability while ensuring that they are developed and implemented in a way that protects and promotes rights. These questions, at the same time, do not exclude the fact that we, as a society, should put into question what are the technologies that we want to represent our possible futures and what are the ones that we want to reject due to their capacity of reproducing and reinforcing discriminatory power (SILVA, 2022).

### 1.2. Preliminary remarks on AI accountability and transparency

AI technologies are primarily complex, opaque, modifiable through updates, capable of learning during operation, frequently unpredictable and vulnerable to cybersecurity threats. Due to these characteristics, the chance of harm posed by these tools is high, and figuring out ways to regulate them has been a challenging task for specialists and policymakers worldwide (PASQUALE, 2015; EUROPEAN COMMISSION, 2020a; BUITEN, 2019).

One of the main challenges posed when considering AI regulation relates to rendering those responsible for developing, selling or operating these systems accountable for their damages for two main, non-exhaustive reasons.

First, determining who is liable for an AI's damages is frequently tricky. The chain of organisations and individuals that make part of the development and deployment of these systems is usually formed by multiple nodes that interfere differently with the technology in distinct stages of its life cycle. Such characteristics may make it more challenging to offer victims of damages caused by these applications a claim for compensation and render individuals and organisations accountable for rights violations (EUROPEAN COMMISSION, 2019).

Second, because many AI systems, particularly those based on deep-learning (such as those found in face recognition or voice assistant applications), display functions so complex that sometimes not even their creators can understand. The databases through which they train and learn are usually so vast, and the chain of decision-making between the different nodes of these systems is so complex that it is usually barely impossible to understand how they have reached a specific decision. As such, despite the far-reaching presence of AI systems, the complexity of their internal operations makes few of these systems intelligible to humans, which makes it challenging to understand the real impact they have on society and how to hold those responsible for them accountable (BAYAMLIOGLU, 2018).

Accountability is a word that has been subject to many different definitions and lacks direct translation to some languages, such as Portuguese. However, one can consider it as a form of relationship in which an agent must explain and justify his or her conduct to another actor or a larger audience, and for such conduct this agent can be

questioned, judged and face the consequences (BOVENS, p. 450, 2017). In an environment concerning the uses mentioned above of AI, we would consider the agent the individual or entity responsible for the system, the ones questioning him or her a regulator or society as a whole, and the conduct would be the pitfall of the system which might be leading to rights infringements.

To address the challenge of rendering these agents accountable despite the complexity of AI, many have argued that promoting transparency is a fundamental step (RUDIN, 2019; MALGIERI, 2018; MITTELSTADT, RUSSEL & WACHTER, 2019). The logic is that once we are capable of understanding how the decision-making processes of AI systems work, regulators and society would be more capable of rendering the agents responsible for these technologies accountable for the errors they incur.

We may find provisions regarding this theme in different legislations worldwide, but frequently in a generic and abstract manner, without much specification on how to attain such transparency. In Brazil, for instance, the *Lei Geral de Proteção de Dados* (General Data Protection Law, Law n. 13,709/2018, hereinafter "LGPD"), in its Article 20, provides for a right for data subjects to request information on the criteria and proceedings used by an automated decision system to profile him or her based on personal data. A similar provision in the European Union's General Data Protection Regulation (Regulation (EU) 2016/679, hereinafter "GDPR") 's Articles 13(2)(f), 14(2)(g) and 15(1)(h).

Bills aiming at regulating digital platforms in both Brazil and the UE also provide for AI transparency. We may find obligations in that sense in the Brazilian Bill PL 2,630/2020, which aims to enact the *Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet* (Brazilian Law for Freedom, Responsibility and Transparency on the Internet) and the European Union's Digital Services Act (2020/0361 (COD), hereinafter "DSA"). Similar examples are the Brazilian bill *Marco Legal da*

*Inteligência Artificial* (Artificial Intelligence Legal Framework, PL 21/2020)[1] and at the EU's Artificial Intelligence Regulation proposal (2021/0106). However, all of them also make reference to generic terms such as information about "criteria", "proceedings" and "logic" when addressing what should be subject to information obligations regarding AI systems.

The lack of clarity of such legal acts does not give regulators much light on how to put AI transparency in place, and even whether or not there is a right to transparency or explanations regarding data processing through automated decision making systems. The challenge becomes greater when we realise that enhancing transparency inevitably leads us to a myriad of barriers that rise from both systems' inner functioning and the social, economic and political contexts in which these applications are developed and applied. To make reference to only a few of such obstacles, one can mention the intrinsic complexity of AI — especially of deep learning systems, as mentioned earlier —, the commercial secrets surrounding them, as well as the possibility of manipulating and gaming algorithms, be it by internal or external agents to their operators (VEALE, 2017).

Apart from this, transparency in itself can bring issues that may render it useless or even prejudicial. Ananny and Crawford (2018) have argued, for instance, that transparency can create even more opacity when an organisation discloses so much information that one cannot understand what is important for accountability purposes and what is not. This is what they call "strategic opacity" or "resistant transparency".

---

[1] After criticism from civil society organisations and specialists, the Rapporteur for the Senate of the *Marco Legal para a Inteligência Artificial*, Senator Eduardo Gomes, commissioned a Commission of Jurists to prepare a new draft for the *Marco Legal para a Inteligência Artificial* on the 30th of March, 2022. The bill's scope is expected to considerably change after the presentation of the new text. See, e.g., SENADO. Comissão de juristas da inteligência artificial faz balanço de audiências públicas. **Agência Senado**, Brasília, 16 May 2022. Available at https://www12.senado.leg.br/noticias/materias/2022/05/16/comissao-de-juristas-da-inteligencia-artificial-faz-balanco-de-audiencias-publicas. Last access: 12 July 2022. LEMOS, Alessandra et al. **Nota Técnica PL n. 21/2020**: sobre o marco legal do desenvolvimento e uso da inteligência artificial no brasil. LAPIN, Brasília, v. 1, n. 1, p. 1-49, nov. 2021. Available at: https://lapin.org.br/2021/11/09/nota-tecnica-atualizada-discute-o-pl-21-a-2020-do-marco-legal-de-ia/. Last access: 12 jul. 2022; DE PEREIRA, José Renato Laranjeira de; MORAES, Thiago Guimarães. Promoting irresponsible AI: lessons from a Brazilian bill. **Heinrich Böll Stiftung,** Brussels, 14 Feb. 2022. Available at https://eu.boell.org/en/2022/02/14/promoting-irresponsible-ai-lessons-brazilian-bill. Last access: 12 Jul. 2022.

In this sense, some of the inevitable questions we have to raise to start any discussion regarding artificial intelligence transparency relates to why, how, and to whom is transparency and accountability relevant. After all, as we will further discuss, any analysis of algorithms cannot be made separately from their social context (BUCHER, 2018), in order to understand these systems not just as code and data, but as "assemblages of human and non-human actors" (ANANNY & CRAWFORD, 2018, p. 983). It is fundamental to understand what for and to whom accountability should be driven, and therefore ask ourselves, "what is being looked at, what good comes from seeing it, and what are we not able to see?" (ANANNY & CRAWFORD, 2018, p. 985).

This dissertation aims to shed light on the opacity of AI-based technologies, or what some have called the "black box" of AI (PASQUALE, 2015), with a focus on how to regulate it. Since AI is a family of technologies, we anticipate that our main focus will be machine learning (ML), a gender of technologies pertaining to the family of artificial intelligence.

In this sense, some of the questions we aim to answer in this work are:

1. What makes a machine learning system opaque?
2. Is opacity a problem?
3. Can transparency be an effective tool to address such problems?
4. Are data protection rules suitable to address machine learning transparency?
5. Can regulatory theory help enhance machine learning transparency by enforcing data protection rules in Brazil?

We expect that these questions will allow us to answer the main research question of this dissertation: *can the responsive regulation theory enhance machine learning transparency in Brazil by enforcing data protection rules prescribed by the LGPD?*

My choice for the responsive regulation theory lies in its ability to design regulation that is adaptable to the reality of each regulated entity in order to grasp the specificities of its activity and behaviour. As such, this work agrees with the theory that there

are no optimal or best regulatory solutions, as practical approaches may vary on the specific market, the historical context and the businesses involved (AYRES & BRAITHWAITE, 1992, p. 5). As a machine learning system may pose different degrees of risk according to its application, we need a theory that does not rely on a one size fits all approach but is highly adaptable.

The regulator's role in responsive regulation is to be attentive to different businesses' and markets' characteristics, identify when and how to take action, and design norms most suitable to different realities. In this sense, regulatory objectives are, according to Ayres and Braithwaite (1992, p. 6), more easily achieved "when agencies display both a hierarchy of sanctions and a hierarchy of regulatory strategies of varying degrees of interventionism". State intervention in businesses escalates and de-escalates according to the level of compliance of regulated entities by applying the so-called "regulatory pyramids", as we will describe in more detail further in this dissertation.

The path we will follow through in our adventure is as follows. In **Chapter 2**, we will define the concepts we will use throughout this work and discuss how data and algorithms interact for ML models to issue outputs. The Chapter will also dive deeper into how many of these systems have been developed in a way that prevents users and regulators from understanding how they work. Finally, we will investigate how these machines have affected society's functioning and why their transparency matters.

**Chapter 3** focuses on what it means for ML to be transparent. We will (**i**) investigate the different taxonomy related to providing and assessing information of these systems, such as interpretability, explainability, justifiability, etc.; **(ii)** have a glimpse of how explanations about ML systems can depend on the criteria such as moment, scope and degree; **(iii)** assess the limitations of transparency in allowing individuals, groups and the state to take action against abuses carried out with these technologies; and then (**iv**) analyse how legislation addresses this theme. We will pay special attention to the means for regulators to be more straightforward in their demands on what and how information should be displayed regarding AI systems while assessing the obstacles for transparency to allow for effective AI accountability. At the end of the Chapter, this work will present a set of questions to be asked by anyone aiming to un-

derstand a machine learning system in order to define what information is necessary and in what format to enhance comprehension about it.

In **Chapter 4,** this work will focus on an assessment of data protection regulations from Brazil and the European Union, namely the *Lei Geral de Proteção de Dados* and the General Data Protection Regulation, respectively. We will assess the evolution of data protection rules over time and how they affect the deployment of machine learning systems. Finally, we will address the open questions for dealing with the risks left open in data protection laws.

Finally, **Chapter 5** will investigate the adequateness of responsive regulation's participatory strategies and the incentives framework it provides for making ML systems more transparent. That includes assessing issues such as what factors may influence a regulator in enforcing rules towards ML systems, the role of enforcement pyramids, and what specificities the theory should have in countries of the Global South.

Machine learning has created great opportunities for humanity to solve long-dated problems. From supporting cancer treatments to allowing for knowledge to reach larger audiences through automated translation applications, ML's potential to support society is significant. However, the risks these systems pose are as vast as their benefits, and the world needs to address them in a way that does not underestimate the complexity surrounding these machines. And we need to understand ML better to profit from its promises in a way that protects our rights.

This work expects to bring inputs for this discussion from a regulatory perspective. After following this journey, it expects to have conceived tools for supporting regulators and regulated entities in defining how to provide information about ML applications that is effective in enhancing accountability and understanding of how these systems affect our world.

# 2. MACHINE LEARNING AND OPACITY

## 2.1. Machine learning definition and methods of learning

A theoretical glimpse on how a machine learning (ML) system works is crucial to identify how automated predictions and decisions that influence our lives on a daily basis are made. Therefore, the present chapter will discuss how data and algorithms relate in order for ML models to issue decisions. It will also address how most of these systems, and the rationale regarding artificial intelligence and machine learning as a whole, have been developed in such a manner that prevents users and regulators to understand how these machines work and how they reproduce power dynamics.

Defining AI has been a huge challenge for experts. According to Russel and Norvig (2013), it refers to the field of study and development of rational agents, i.e., entities that perform actions, acting on an environment, in order to achieve the best possible expected result.

Another definition, which tries to cover also the proceedings through which AI learns and acts in the world, is the one provided by the European Union's High Level Expert Group on Artificial Intelligence (2019). Similarly to Russel and Norvig's considerations, it highlights how AI is both software (and sometimes a hardware) development and a scientific discipline. On the other, it is descriptive on how AI systems are designed by humans to "act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal".

Machine learning (ML), on its turn, is a branch of artificial intelligence, and has become one of the most common technologies applied for tasks that require information extraction from large datasets, and has been used in systems such as search engines, social media feeds, credit scoring systems and predictive justice applications. It refers to a set of techniques for building models capable of automatically detecting, discerning and operationalising meaningful patterns in data without explicitly programming them, by inducing their development from specific datapoints. Each technique uses a

different approach to extract and encode such patterns in a way that they are likely to generalise and then applied to new information that is not present in the data used to train the system (SHALEV-SHWARTZ & BEN-DAVID, 2014; VEALE, 2019).

ML systems "learn" to perform specific activities by extracting information from a specific, pre-existent, dataset, which we can call *training dataset*. Having learned from information contained in such a pool of information, they can process new data (input) so as to perform a task (output), which can be a classification or a prediction, for instance.

One common way to categorise different approaches to teach these machines relates to the degree of human supervision involved in its learning process. A form of classification is between supervised, unsupervised and reinforced learning.

A system that learns through **supervised learning** gains experience by analysing a dataset already containing significant information that will allow the model to perform its tasks (SHALEV-SHWARTZ and BEN-DAVID, 2014). A common example is an email spam filter. Based on past experience acquired by processing information on what sort of patterns are usually followed by spam messages, machine learning systems are able to identify what sort of information in a communication are most likely to characterise it as spam. Spam filters usually learn from pre-classified training data that contain samples that have already been categorised as spam or as legitimate messages so as to grasp what are the characteristics most common to them and thus learn how to carry out classifications more or less independently of human review. Based on this experience, and also by analysing, for instance, datasets containing lists of common spam senders, the ML application can analyse a new message (input) and later classify it as spam or as a legitimate email (output) (TRETYAKOV, 2004). The fact that the training data is already labeled when the system starts to learn from the dataset makes it a form of supervised learning algorithm.

In **unsupervised learning**, on the other hand, there is no distinction between training and test data. It allows to approach a problem with little or no idea on what results should look like, in a way that there's usually no feedback based on the prediction results. Clustering a data set into subsets of similar objets based on relationships among

the variables present in the data (SHALEV-SHWARTZ and BEN-DAVID, 2014; NG, 2022) is a typical example of application that can be trained with the use of unsupervised learning. They can be applied, for instance, to separate users of a social network into personality categories based on their profile information with the aim to target advertising more effectively for each different group of users. Besides clustering, other unsupervised learning systems include, e.g., association rules, frequently deployed in recommendation systems in e-commerce websites or streaming services (PORTUGAL; ALENCAR; COWAN, 2018).

Supervised and unsupervised learning are the two extremes of a broad spectrum of supervision which is occupied by machine-learning systems. Most frequently, systems will occupy an intermediary level of supervision (GOOGLE DEVELOPERS, 2022). That's why some authors make reference to **semi-supervised learning** algorithms, found mainly when working with a training set with missing information, where only a part, though a significant one, of the dataset is properly classified (CHAPELLE, SCHÖLKOPF, ZIEN, 2006). In this sense, instead of discarding unclassified data, certain pieces of information about them can be used to help improve the model (VEALE, 2019). Examples can be found in movie rating websites, where systems are responsible for grading movies in a context where not every user has rated every movie (PORTUGAL, ALENCAR, COWAN, 2018).

Finally, it is also meaningful to mention another approach for teaching these machines: **reinforced learning**. Algorithms working under this umbrella learn "based on external feedback given either by a thinking entity, or the environment" (PORTUGAL, ALENCAR, COWAN, 2018, p. 4). This is usually the approach adopted for ML systems that play board games against opponents. Good moves receive positive feedback, while bad ones are discouraged. With the feedback received, systems are able to learn how to best perform in these environments.

The finite and sequential set of computational steps that transform input into output for solving a specific problem consist of the so-called *algorithm* (CORMEN; LEISERSON; RIVEST; STEIN, 2009), as it is understood under what Veale (2019, p. 27) calls "repeatable recipe definition". For him, however, this somewhat cold, senseless

definition of algorithms falls short of comprehending the fact that algorithms are always relating to their contexts while reproducing social values, and never only a mere sort of technical rules neutrally arranged.

By combining the definition we provided for machine learning with a description of the different approaches used for teaching ML systems to perform actions, we can also understand the centrality of data for these applications to operate. Much of the effectivity of a system involves how it develops itself through the processing of information it accessed on its environment. For this reason, it is fundamental not only to understand the technicalities around data processing, but also to assess critically how data is usually collected, accessed, categorised and applied for training machines. The role of humans in this process is pivotal, and the way they plan, design, sell, comment and operate these systems has key implications for the welfare of our society and our planet.

### 2.2. Machine learning and data

The aforementioned definition of machine learning systems as being capable of detecting automatically meaningful patterns in data brings to light the centrality of datasets for the development and operation of these systems. As such, their performance will also be directly related to the quality and the amount of data gathered in the training dataset that was applied for the learning of these systems.

Along with other kinds of artificial intelligence applications, machine learning saw an outstanding peak of development with the evolution of automated data-gathering techniques and the huge price reductions of mass-memory storage and processing apparatus that have allowed for the collection and retention of incomparable amounts of data. This was fundamental for the technology to improve. Machine learning systems had vast datasets to learn from and build capacity to process even more data, and thus keep classifying, quantifying, and extracting useful information from them (NILSSON, 2009).

In this scenario, a common mindset came to permeate the mindset of most of the technical community developing these systems. The motto "the more data the better" to feed ML systems came to be taken as something *a priori*, as machine learning applications were, at least apparently, enhancing in exponential scale in comparison to what happened in previous decades due to the greater availability of data (NILSSON, 2009).

Machine learning, hence, has taken advantage of and also perpetuated what came to be known as big data, a term that has been mostly used to refer to extremely large datasets that could first only be captured, managed and processed by high performance computers, and that now can be, to some extent, processed also by standard softwares (MANOVICH, 2011).

Even though a phenomenon originated a few decades ago, big data became a buzzing word and an object of frenzy in more recent years. As Michael Veale (2019, pp. 25-6) puts it, this "allure of big data" refers not only to the technical phenomenon around the use of vast amounts of datasets to extract predictions from them, but also a social, cultural and mythological tool developed by stakeholders in the industry to feed the interest on the models and insights possible to extract from these information pools. The success of machine learning applications in paradigmatic cases such as the victory of a system developed by Google subsidiary Deepmind Alpha Go against the professional Go player Lee Sedol made it seem like these systems were capable of anything as soon as they were built on the foundation of mass data processing.

In this sense, besides being a very large pool of information, big data should not be taken into consideration apart of the tools that extract relevant patterns from data, especially machine learning (HILDEBRANDT, 2013). As we have mentioned, machine learning was mainly enhanced because of the vast amount of data from which it was fed, and big data started to become more and more useful as the techniques for extracting meaningful information from it became more effective, in sort of a feedback loop.

As such, this massive use of data to train machine learning systems became a precept in the development of machine learning systems to the extent that one should not care

about the messiness of information found on datasets, as long as the amount of available data for use was the greatest possible.

That's why scholars such as Mayer-Schönberger and Cukier (2013) came to vocalise that "more [data] trumps better", meaning that the quality of datasets should be a marginal worry for anyone developing machine learning systems. This notion would also be behind the annunciation of an era approaching in which sampling information for taking decisions will be unnecessary, as we now have access to all the necessary data and that using it would allow humans to see details which would be impossible to assess in smaller quantities of data. In this sense, the authors posit that *N=all,* i.e., the information that we need is already there for us to access, and samples are no longer necessary to deduct information from them as big data is rapidly approaching a point in which it will give us a full picture of society. Important to note that this optimistic position has been overly criticised by scholars such as Hildebrandt (2013), Crawford (2021) and Veale (2019), as we will show below.

By stressing this, Mayer-Schönberger and Cukier also made reference to what Halevy, Norvig and Pereira (2009, p. 9) meant, while discussing the development of natural language processing (NLP) systems, when they affirmed that "invariably, simple models and a lot of data trump more elaborate models based on less data". It is important to note that NLP is not machine learning, but stands by its side as a subset of AI (KHANBHAI et al., 2021).

NLP models are a good example of how data quantity was fundamental for their enhancement, which happened mostly after developers started to train them not through the enunciation of complex linguistic formulas, but through the assessment of enormous amounts of data in datasets containing as much as billions or trillions of words found on the internet. For this reason, Halevy, Norvig and Pereira (2009, p. 12) affirmed that web would be an incredible source of data, where a "large training set of the input-output behavior that we seek to automate is available to us *in the wild"*. That led the authors to issue a call that also reflects much of what has been taken for granted among enthusiasts of machine learning: "[n]ow go out and gather some data, and see what it can do".

These transformations were held against a background of a massive datafication of the world, which refers to the technological trend turning many aspects of our life into data and its subsequent transformation into information through the application of software to process and analyse such raw data to realise it as a new form of value (MAYER-SCHÖNBERGER, CUKIER, 2013; ZUBOFF, 2019). In this process, not only human creations and communications have been transformed into machine-readable information, but our own bodies and minds became objects that many organisations have been trying to access to train their systems and fulfil their interests, frequently to the detriment of groups and individuals affected by them, in a process called by Zuboff (2018) as "rendition".

That's why critics have been increasingly vocal on how ideas such as the ones expressed above by both Nilsson (2019) and Halevy, Nerving and Pereira (2009) are dangerous. Not only for every quantification there is an unavoidable preceding qualification, which leads us to question who is responsible for such a preliminary data assessment and how it is carried out (CRAWFORD, 2021), but also that using all the data available from specific sources "does not reflect the necessary incompleteness of knowing the world in a quantified way" (VEALE, 2019, p. 37). *N=all* is, hence, a misleading idea, as the translation of life into data can be made in infinite ways according to how existence is experienced by each individual or group. Whichever way this process is made has a "major impact on the outcome of the data mining operations (HILDEBRANDT, 2013, p. 32).

Further, affirming that data is out there, "in the wild", for anyone to catch, also means that someone might not be much worried about what are the stories and narratives behind these information and the individuals and groups to whom they refer. This opens a series of ethical and legal questions regarding privacy, data protection and inequality and many others that put into question the idea that machine learning systems are neutral, objective tools ready to interpret the world.

It is also important to underline that, although machine learning systems are able to trace *correlations* between pieces of information, they are not capable of establishing relationships of *causation* between them. They can present a prediction or e.g. iden-

tify a face in a picture, but not to establish how or why they got to this conclusion. They might respond to “what” questions, but not “why” (MAYER-SCHÖNBERGER, CUKIER, 2013). As these systems are being deployed by governments and industries worldwide to issue decisions as critical as whether we can have access to credit or if we look like someone wanted by a law enforcement agency, the difficulty to assess how and why these systems issue decisions is a major threat to our fundamental rights.

To exemplify how the quality of the data training an algorithm affects its functioning, we can assess the building of a computer vision system that differentiates apples from oranges in photographs.

In order to develop such a system, the first step would be to organise a database with plenty of pictures of apples and oranges. In a supervised learning approach, each image will be labeled as that of an apple or of an orange. The algorithm will thus survey these data to develop a model for identifying the difference between the two classes of objects and thus be able to distinguish the difference between them (CRAWFORD, 2021).

However, let’s say that our developer, when building the training dataset with the images of apples and oranges, only included pictures of red apples, forgetting completely to add green apples to the database. In this situation, the system will automatically conclude that every apple is red, having no idea that the poor green apples should also be categorised as such. (CRAWFORD, 2021)

This simple situation demonstrates the capacity that most machine learning systems have to extract hypotheses from data. They make inferences based on the information which they have access to, and not necessarily on logical premises. For this reason, the system in the aforementioned example could not recognise green apples as a part of a broader category of apples along with their red counterparts. That’s why machine learning systems’ inferences are inductive, rather than deductive. Inductive inferences are mere hypotheses based on previous information, while deductive are conclusions following necessarily and logically from their premisses. For this reason, inductive

references are always subject to change by assessing additional data (NILSSON, 2009, p. 495).

Maybe failing to differentiate apples from oranges would not be much of a burden to society. But if we extend this analysis to a type of technology such as face recognition, which is being increasingly adopted by stakeholders ranging from intelligence and law enforcement agencies to advertising companies and financial institutions, the impacts to society can be much more problematic.

Face recognition technology (FRT) may be used for different purposes. It can be applied at least to four goals: to identify individuals by determining to whom belong a specific face; to authenticate, by defining whether a person is indeed who he or she claims to be; to (purportedly) identify emotions based on the contraction of muscles in someone's faces; or to define the gender of an individual (MORAES; ALMEIDA; PEREIRA, 2021).

Similarly to our system differentiating apples from oranges, to build a face recognition application one needs a broad dataset with images, but now of human faces. With that in hand, the developer will then trace a set of rules for the algorithm to scan unique identifying details of a person's face so as to differentiate her images correctly from all of the other individuals in the database. The system does so by measuring, for instance, the distance between the eyes of a person and the distance from her forehead to her chin. The result is known as the 'individual signature'. The algorithm compares this signature, which is a sort of mathematical formula, to known faces in a database (MORAES; ALMEIDA; PEREIRA, 2021). Based on the similarity of the model captured and the data found, a match between the image captured by the surveillance camera and a given image in the faces database may be made.

As the use of face recognition systems became far more frequent, the errors they incurred were soon seen as troublesome. They have been failing to identify properly individuals, allowing even for the detainment of innocent individuals based on the mistakes of these systems. In Brazil, for instance, a woman was imprisoned after being misidentified by an FRT with a wanted woman that was already imprisoned. That happened because the database that the system was using to identify individuals was

not updated (WERNECK, 2019), something which draws attention to the lack of attention given to the quality of the data feeding the system. A similar situation happened in the US, where a man was mistakenly detained after being misidentified by an FRT (HILL, 2020).

Face recognition has also been responsible for targeting far more errors when assessing face images of individuals with darker skin tones than the opposite. It came to be proven that these frequent mistakes were happening mainly due to the lack of representativity of non-white individuals, especially women, in the databases used for training the applications (BUOLAMWINI; GEBRU, 2018).

These expressions of race and gender biases in face recognition systems are pivotal examples of how machine learning systems are far from neutral. They are unavoidably embedded with the categories, the data, the biases and the interests of those developing them. As such, any regulation for addressing the risks emerging from these applications should start from the idea that neutrality is unreachable when taking into consideration the creation of any human tool, especially one that depends on the analysis of historical data, something already unavoidably inlaid with bias.

As Kate Crawford (2021) puts it, the practices of classifying information are inherently political. In the field of machine learning, they stigmatise and reduce individuals and their experiences to categories so that they can fit the limitations of machines. When developers or crowdsourcing workers categorise individuals based on their images for carrying out assessments of their skin colour or their emotions, for instance, they are exerting power in how these people will be interpreted by the systems they are feeding. As the outputs of ML shape reality based on these classifications, they do so based on the interests and biases of the very individuals that participated in the creation of these machines.

That's why attempts to de-bias datasets such as the one carried out by IBM (MERLER *et al.*, 2019) in their face recognition systems had foundational problems. By trying to diversity their datasets with more images of women and darker-skinned individuals, they did not take into consideration that the very classification of male and female left apart a series of individuals that do not see themselves represented by the

reductionist binary gender classification. And the same applies to skin colour, as categories such as race, ethnicity, culture and geography go far beyond features in our faces, but encompass political, social and cultural backgrounds that cannot be captured by a camera (CRAWFORD, 2021).

After all, it is necessary to have in mind that "all information systems are necessarily suffused with ethical and political values", especially those that have disappeared "into infrastructure, into habit, into the taken for granted" as affirmed by Bowker and Star (1999, p. 321). That also extends into the datasets used in machine learning systems, fully embedded by the economic, social and political interests and biases of the ones developing and operating them.

Data quality of datasets, thus, despite being an important step to develop more responsible ML systems, is not sufficient per se. We must ask ourselves why these systems exist, what sort of data do they use along with the human stories they hide, and why are they embedded with the specific steps assigned in the algorithms that shape them. The example of face recognition is always representative of how structural racism is present in society, and the history of how they were built with the use of categorisation techniques based in racist stereotypes makes us question why this sort of technology still exists (CRAWFORD, 2021; SILVA, 2022; ALMEIDA, 2019). Reducing its flaws to merely technical aspects is thus misleading.

However, the answers for these questions are hardly answered. The opacity regarding the datasets used, the main rules and logic governing in their algorithms, and how specific decisions affecting people's rights were made became a common place in machine learning, along with the lack of justifiability on the motivations behind the deployment of these systems and to what extent their outputs are legally acceptable (MALGIERI, 2020).

Pasquale (2015) calls this a feature of the "black box society" in which we are trapped. In the black box society, information is detained by a few organisations and governments and kept under locks and keys even from the persons affected by these algorithmic systems, on the basis of justifications such as governmental confidentiality or commercial secrecy. ML systems have thus become through the years more and

more wrapped in a veil of mystery due to their technical complexity, which also serves as a convenient excuse for guarding the secrets of the groups that develop and deploy them.

As a response to such opacity, much has been debated regarding the transparency of machine learning applications. Nonetheless, questions such as which systems should be explained, what is the level of explainability to be provided, as well as what explainability actually means have not yet been answered (ROBBINS, 2019). We also cannot escape asking ourselves whether explanations are useful at all (ANANNY, CRAWFORD, 2018; EDWARDS, VEALE, 2017). Meanwhile, systems continue to be applied to inform high-stakes decisions and the accountability for their errors and injustices keeps lagging behind due to the lack of relevant information regarding them for regulators.

### 2.3. Opacity: beyond the black box metaphor

As earlier mentioned, ML systems have been frequently described as black boxes, in the sense that their inner functioning would be unintelligible by those making use of them, which sometimes include even their own developers. This lack of transparency is usually linked with the frequent lack of accountability for unfair or illegal outcomes incurred by different ML models (MOHSENI, ZAREI, RAGAN, 2018). This is based on the idea that, without being able to understand them, judges and regulators cannot properly address who should be liable for a negative outcome of a system, and groups and individuals affected by it would hardly know how their lives are being influenced by their operations. Another obstacle in pushing for further ML transparency is the challenge of balancing human rights with trade secrecy protections detained by companies deploying or making use of these systems, a feature that poses further barriers for substantive intelligibility.

This exact discussion has happened in a recent decision issued by the First Regional Labour Court in Brazil in a lawsuit involving a driver who sued Uber for the recognition of employment relationship. The plaintiff, among other demands, requested that

the Court ordered an audit of the company's algorithm to assess to what extent the ML system used by Uber created a structure of labour subordination between the driver and the app. Despite Uber's outcry that assessing its algorithms would violate trade secret prerogatives while also putting in risk the company's business model, the Court decided that the auditing would be reasonable as it would be the only way to assess to what extent the driver was subordinated to the commands of Uber's ML systems (BRASIL, 2021).

Even though the Brazilian decision was not final, it is an interesting example of how the tension between the demand for accountability, trade secrecy protection and whether assessing the code of an algorithm is useful or not is put when analysing how human beings are affected by ML applications in their fundamental rights. Further, it represents how companies spend a great amount of their capital and power in trying to keep information on their systems under a veil of opaqueness, involved by what Pasquale (2015) calls a black box. All of that despite the fact that these technologies have been built over the knowledge produced by decades of public-funded research (PIKETTY, 2020).

The black box metaphor has been widely used to refer to the difficulty in understanding the inner workings of algorithms. However, scholars such as Bucher (2018) have been putting into question the convenience of this metaphor. According to her, defining algorithms as black boxes might sometimes consist of a sort of *strategic unknown*, whereby organisations have been comfortably avoiding too much of an effort to render their systems intelligible in order to avoid accountability for their outcomes.

Arguments frequently used to avoid this, so to say, opening of the black box, are that releasing information on algorithms may lead bad-faith actors to game the systems and that such information is legally protected as a trade secret (BAYAMLIOGLU, 2018). Another statement made by developers and organisations is that it is impossible to completely understand ML applications due to their complexity, which make them opaque even to their own developers. They do not, however, affirm why should we be relying on technologies that are so easy to be gamed (VEALE, EDWARDS, 2017).

To some degree, this is not exactly wrong. ML systems in social media, for instance, are usually works of collective authorship, made, maintained, and revised by many people with different goals at different times. Further, as many systems are built on top of others, and work in conjunction with other algorithms constantly receiving new data to predict and train, there is a point in which they reach a certain level of complexity that their outputs can be difficult to predict precisely (SEAVER, 2013). Nevertheless, as Bucher (2018, p. 57) puts it, this is an expression of what McGoey calls knowledge alibis, which is "the ability to defend one's ignorance by mobilizing the ignorance of higher-placed experts" (MCGOEY, 2012, pp. 563-4). The idea is that, when even experts cannot understand the details of algorithms, how could companies be able to do it?

Despite the potential for opacity being used as a convenient excuse, Bucher (2018) calls attention to the fact that one problem with the black box metaphor is the fact that it presupposes that all one needs is to go there and open it in order to understand its inner functioning. This might not be always useful, and may be even misleading.

As we noted, algorithms are currently deployed in complex chains in which one feeds the other with the inferences it reached by assessing a given dataset. In such cases, understanding each and every algorithm, then the functioning of a whole ML apparatus and finally how every outcome is carried out, might be indeed almost impossible to pursue, and might not even be desirable. After all, most of the time it is not exactly the source code of an algorithm that we need to comprehend why and how an injustice is being done. Instead, sometimes it might be more useful, for instance, to understand what are the political, social, environmental and economic reasons behind both the practices that help sustain the notion of algorithms as black boxes (BUCHER, 2018, p. 59) and the ones that define which, why and how important decisions, that directly affect the lives of human beings, should be carried out by these opaque machines (CRAWFORD, 2021).

Against this background, ML continues to underpin many contestable predictions and decisions. Rendering these systems explainable has thus continued to be regarded as an important instrument for allowing regulators and users to better understand how

these automated decision-making models reach specific predictions and how to revise them in case of mistakes, allowing both developers and users to gain greater control over these systems. However, defining what can be comprehended in these models and, based on that, what is it that one needs to know, for what purposes, and who exactly needs such an information is of fundamental importance to establish what it is that we are demanding when we talk about ML transparency.

Adding to this, Arya et al. (2019, p. 1) have identified that there is a gap between what the research community is producing about ML transparency and what regulators and society as a whole have demanded from them. One reason for this gap is the lack of a precise definition of how these explanations should be carried out, something which is due especially to the fact that "different people in different settings may require different kinds of explanations".

And transparency does not suffice when related only to the technicalities of an ML system itself. Ananny and Crawford (2016, p. 974) suggest that, instead of limiting oneself to merely looking inside them, it can be more useful to look across them, by "seeing them as sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans". After all, as Veale (2019) puts it, algorithms cannot be divorced from the contexts in which they are applied. This means that they are always interwoven with an assemblage of other actors, humans or not, living beings or machines, both in the present and in the past, that are constantly shaping their functioning. As such, they represent a complex network filled with multiple economic, political and social interests that are being embedded in systems that deeply affect our lives, and this information can also be of great importance on assessing whether the use of an algorithmic application respects fundamental rights or not. We will look at this further in the next chapter.

Such a complex scenario poses a major challenge for regulators. From a data protection perspective, for instance, how should one assess whether the developer or deployer of a machine learning system is providing for sufficient intelligibility and is thus fulfilling its transparency obligations on personal data processing and thus allow

for the data subject to exercise her rights under the Brazilian General Data Protection Law (LGPD)? From a more broad, fundamental rights perspective, how should one understand how the processing of data by an ML system might be affecting one's access to a job position or credit due to hidden discriminatory biases based on race and gender, and thus potentially violating the right to equality enshrined in the Brazilian Constitution?

Regulating such a broad set of different applications under the umbrella of machine learning comprehends the need to assess systems being used to different social and economic sectors. Further, the same kind of system may be applied differently in different social systems, which means that a risk-based approach should take into consideration, among other factors, both the sort of application being deployed, the field and the manner in which it is applied, and who it might be impacting (EUROPEAN COMMISSION, 2020b). Returning to our face recognition example, such a system can be used, for instance, to unlock a smartphone or to pursue individuals escaping from the police. In this sense, the myriad of different ML applications, and the different ways in which they are applied puts the issue that different regulators might thus have different parameters on what would be sufficient information about a given ML system.

With that in mind, perhaps the biggest challenge relates to determining what should or should not be explained, since providing excessive information or even transmitting it in an inadequate manner would render the information about a given ML system ineffective and unnecessarily costly (BAYAMLIOGLU, 2018). Hence, identifying how and to what extent should an application and the broader environment in which it is applied be understandable requires thus an attentive eye of the regulator, who will need to assess each family of applications and decide what sort of information should be provided in a given case.

The next chapter aims thus to address what has been discussed in the field of machine learning transparency as a preliminary step to frame how this theme is being addressed by different legal provisions. However, before moving forward to it, it would

be interesting to assess another subject which has been left in the corner when discussing ML's opacity: its labor and environmental impacts.

### 2.4. The back-end of how machines learn: a brief note on the labour and environmental black box

Although the effort of those building and selling machine learning systems in portraying them as something almost magical, this vision underscores what Irani (2016, p. 36) calls the "hidden layers of human data work" that "calibrate algorithms to culture". In this section, we will briefly cover how ML systems are not that autonomous by knowing who are the human beings working to fill the gaps when ML systems are incapable of performing a task and how they are conveniently kept hidden behind the narratives of magical machine intelligence and autonomy. Afterwards, we will also discuss how nature is also affected by the race for ML development, as demand for natural resources increases among a powerful data hungry community.

As we saw, machine learning technology is developed through feeding algorithms with large pools of data. However, they frequently fail in conducting precise predictions for not being able to assess specific nuances in data that require a deeper, sensitive knowledge of the world. This may happen for several reasons, which include the messiness of datasets and the classification problems they incur, as well as cultural specificities with which systems do not know how to interpret due to the lack of data representativeness of a specific population. This is quite common when one needs to solve issues related to content moderation tasks such as differentiating hate speech from sarcasm, or pornography from artistic nudity. In these situations, developers frequently need to go beyond feeding ML with even more data to solve them (GRAY, SURI, 2019).

That's where low-waged human work gets into the equation. To fill in the incapacities of ML systems, companies have been delegating tasks that cannot be carried out by these models alone to human beings. These people are responsible to help feed machines with information that they would not be able to identify for themselves or even

to fix the errors in which they incur. One example is the hiring of content moderators by social media companies to identify nuances in speech and the reliability of information before flagging content as misinformation. Other is the use of crowdsourcing platforms to have groups of people tagging objects in images that will serve as training data for image recognition systems (CRAWFORD, 2021).

This crowd of "ghost workers", as Gray and Sury (2019) refer to them, is responsible for doing the necessary job to keep ML algorithms performing functions related to the interpretation of cultural data that they are not able to process on their own. They are allocated in projects to perform micro functions such as identifying whether a person is the same in two different photographs and watching violence videos online to flag and keep them out of social media platforms like Facebook.

All of this in exchange for a wage of only USD 2/hour, on average, (HARA *et al.*, 2018) and, for those involved in watching toxic content for hours, a great degree of vicarious trauma and other mental health issues (NEWTON, 2020). After all, as Irani (2016) puts it, the work conditions for these individuals will be governed by the market and its own reasoning of convenience. As these jobs do not require specialised skills, the offer of labour force is frequently high, especially in periods of economic crisis (MORESCHI *et al.*, 2020).

Amazon's crowdsourcing platform, the Mechanical Turk, is a good example of how this market works. The platform allows for developers of ML systems from the most diverse fields to find freelancers to execute tasks either as volunteers or in exchange for money. Developers decide what are the services they need and how much they will pay for them, and contractors have no right for minimum wage or any social benefit. In case an employer does not pay them, Amazon hardly moderates disputes effectively, leaving crowd workers most of the times without assistance (IRANI, 2016). To organise themselves to avoid such forms of injustice, Turkers, as these workers are called, have to rely on forums to exchange information on unreliable employers (YIN *et al*., 2016) or, as has been the case in Brazil, on Whatsapp groups (MORESCHI *et al.*, 2020).

This hidden, poorly paid workforce has been underpinning most of machine learning dreams sold by startups around the world, especially in the Silicon Valley. It is a way of black boxing that "pushes obfuscation into deception" (SADOWSKI, 2018) in a race for power and profit by those wanting to surf the AI hype. After all, it is much more fun and profitable to sell a magic product without having to mention that you use unemployed workforce to feed it.

Apart from the exploration of human beings as a means to further develop machine learning applications, one should not disregard also the environmental effects caused by this industry. These impacts can be represented in at least two main categories, one related to the energy consumption necessary to train an ML model, one related to the mining demands to build the devices in which these technologies will operate, such as smartphones, notebooks and ML-powered autonomous vehicles.

To investigate the carbon dioxide emissions of machine learning models, Strubell *et al*. (2019) assessed the process of development of a single natural language processing system. They found out that, in the process of training and developing it, approximately 660,000 pounds of carbon dioxide were emitted in the atmosphere, roughly the same amount of emissions produced by five cars over the cars' lifetime. That for one single model.

Regarding the mining demands of ML systems, when discussing the footprint of mineral resources necessary to build the equipment for machine learning systems to operate, a didactic example is in the industry of electric cars and autonomous vehicles. An example is Elon Musk's Tesla. To build the rechargeable battery pack for each Tesla Model S electric car, the company needs about 63 kg of lithium (LAMBERT, 2016). As such, it is no surprise that Tesla is the greatest lithium-consumer in the world, estimated to use more than twenty-eight thousand tons of lithium hydroxide annually—half of the planet's total consumption (CRAWFORD 2021).

Few of those involved in the development and selling of machine learning systems would be interested in the disclosure of the human and environmental effects of their models. After all, that would mean putting in doubt the magical myth of AI and mil-

lions of dollars of angel investors' money. For that reason, this kind of inconvenient truth has been left out of the table under a veil of secrecy.

That adds to the notion that ML opacity expresses itself not only through technical obstacles in drawing explanations regarding the inner workings of ML systems, but also on their own degree of autonomy (SADOWSKI, 2018) and the carbon footprint left behind during their development. In reality, the opaque nature of machine learning is also underpinned in economic and political choices that hinder deeper assessments on the social and environmental impacts of these technologies, adding another layer to the black box problem.

The lack of transparency has led scholars such as, among others, Rudin (2019) and Doshi-Velez and Kim (2017) to draw attention to the importance of providing meaningful information about machine learning systems that may have an impact in people's lives as a tool for accountability. At the same time, however, Ananny and Crawford (2018) and Edwards and Veale (2018) have highlighted the limitations of the transparency ideal and how it may end up overburdening individuals and giving rise to other forms of opacity. Transparency, according to them, be seen not as an end in itself, but one of many instruments to allow for the exercise of people's rights against the deployment of harmful ML systems.

With that in mind, the next chapter will assess the strengths and weaknesses of ML transparency and investigate what should be considered as a meaningful information to be accessed not only about these models but also about the environments in which they are being designed and deployed. It will further explore how scholars have been defining terms such as transparency, intelligibility and explainability when applied to machine learning, as well as provide a brief overview of techniques used for making ML more understandable.

# 3. MACHINE LEARNING TRANSPARENCY

In the last chapter, we saw how machine learning systems work and how they reflect the biases and intentions of the human beings developing and deploying them. Not only the strict functioning, but also the environments, contexts, in which machine learning technologies are planned and operated are of fundamental importance for a critical analysis of the social, economic and environmental impacts of tools.

We also discussed the fundamental role that data plays in the development and deployment of these systems, and how it consolidates historical perspectives embedded in society. We highlighted how humans using ML do so under a technocratic narrative that most of the cases serves as a perfect excuse for unjust outputs. Finally, we saw how ML opacity looked like, how the black box metaphor should be used with due caution, and in which manner the impacts of this lack of knowledge also extended beyond the mere technicalities of a system, up to the point that they end up impacting also the planet and labour relations.

Having ML opacity and the ways to go beyond it as the central theme of this work, we now turn towards understanding what ML transparency looks like. As promised beforehand, we will (1) describe what we mean with terms such as transparency, interpretability and explainability, as well as how they are usually applied by the computer science community with regards to the moment, scope and degree of the information provided; (2) understand what are the strengths and weaknesses underlying ML transparency and its effectiveness in addressing ML's impacts; and (3) analyse what are the questions that regulators should ask themselves when demanding that ML systems be rendered more transparent in order to more effectively address their impacts.

## 3.1. ML Transparency's Alphabet Soup

The literature on the provision of information regarding machine learning systems uses different names to refer to the promotion of transparency. Due to the diverse definitions used to describe the terms used to provide information about a model, a brief

elucidation of what this paper means for each designation is fundamental for carrying our discussion forward.

It is important to note that the lack of consensus leads to a myriad of meanings given to terms such as transparency, explainability and interpretability, and cause a considerable confusion among scholars. It is thus hard to find a definition which is authoritative enough to guide this work. For this reason, we will describe how each of the three concepts mentioned above — transparency, explainability and interpretability — is described by selected authors and, from their views, find a meaning that will guide us through the following pages. We will start with *transparency*.

*Transparency*

The term *transparency* has been widely used as a general way to refer to the provision of information regarding ML systems. It is commonly referred to as a means to empower a given stakeholder to trace, explain, communicate and discover information about the working of a system. One can see this idea, for instance, as a principle to be followed by groups involved in the development and operation of not only ML, by AI in general, in different ethical and rights-based approaches to principles for artificial intelligence. These principles-based proposals proliferated in the last years, and have been object of thorough consideration due to their lack of enforceability (MITTELSTADT, 2019).

According to a report aimed at mapping thirty-six different principle-based proposals drafted by stakeholders from different sectors and continents for AI systems, in most cases the principle of "transparency" has been translated as a command that AI systems should be designed and implemented in such a way that oversight of their operations is possible Fjeld et al. (2020).

We refer to two of these ethical guidelines to exemplify how this notion of transparency is addressed. The first is the one drafted by the European Commission-mandated High Level Expert Group on Artificial Intelligence (HLEG, 2019), which argues that transparency should be read as a *requirement* for AI systems to be (quite re-

dundantly) transparent with regard to the data, the system and the business models. Transparency would encompass the notions of traceability, explainability and communication, being thus the gender of which these three concepts would be the species.

*Traceability* would refer to the data sets and processes that inform the AI system's decision, including documentation and record-keeping of data gathering and data labelling techniques as well as of the algorithms used. *Explainability* would concern "the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)", so that a system can be understood and traced by humans. Finally, by *communication* the guidelines mean a right for humans to be informed that they are interacting with an AI system (HLEG, 2019, p. 18).

The other ethical guidelines that we find suitable to refer to are IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2021), which resulted from the works of stakeholders from six continents from academia, industry, civil society, policy and government. According to the study, transparent systems are the ones that allow one to discover how and why a system made a particular decision so as to reduce and magnitude of harm by helping users understand the system they use and helps ensure accountability. They must be transparent to a wide range of stakeholders for different reasons, and for each of them the level of transparency will defer.

Different ways of conceptualising transparency can also be seen among researchers. Rader, Cotter and Choo (2018, p. 2) acknowledge that transparency of algorithms is usually seen both as "the act of making a system knowable or visible" and the "state that is the outcome of a process" of rendering such knowability. Differently, Annany and Crawford (2017, p. 975) affirm that "transparency is thus not simply 'a precise end state in which everything is clear and apparent,' but a system of observing and knowing that promises a form of control".

Based on the aforementioned works, this dissertation sees transparency as the *act and the outcome of making a system knowable, visible, and thus understandable, to an individual or group.* Since this can be made through different ways of providing information about it, we consider transparency *the gender to a myriad of specific techniques (species) that may be deployed by those governing these systems to achieve the*

*desirable comprehension of the given machine learning tool and the environment in which it is developed and deployed.*

In this sense, this work applies the concept of transparency every time it does not refer to specific techniques, but instead to the act of making a system understandable by an observer, to whether its features can be comprehended or not (transparent, not transparent), and the degree of how understandable it is (more transparent, less transparent).

In order to address the species encompassed by the concept of transparency, one may look to two common, different terms widely used by scholars: explainability and interpretability.

*Interpretability, Explainability*

*Interpretability* and *explainability* are additional concepts that also have not yet found their way towards a broader consensus on what they mean. Tim Miller (p. 14, 2018), for instance, equates both terms to refer to "the degree to which an observer can understand the cause of a decision" and define explanation as "one mode in which an observer may obtain understanding, but clearly, there are additional modes that one can adopt, such as making decisions that are inherently easier to understand or via introspection". Other authors follow suit in equating both terms, such as Hamon et al. (2022).

Differently, Cynthia Rudin (2019) posits that there are differences between explainability and interpretability. She argues that explainability would be a form of rendering black boxes comprehensible with external tools and thus have them to issue "explanations". Interpretability, in its turn, would be a quality of systems which are by design intrinsically comprehensible, and thus do not demand external explanations about their mechanisms.

Rudin's perspective thus differentiates:

- a system which by design can have its inner functionings assessable by a human being, either through assessing its code, its training data and the weights given to specific data points, or through comprehending the path the system takes for making each decision. This is what she would call *interpretability;*
- the use of tools external to the ones already part of a by design black box system to understand its features (inputs, rules and outputs) through human-readable information. An example of explainability technique can be, for instance, counterfactual explanations that are capable of showing how a different output could have been reached if any input was changed in a certain system, as Wachter et al. (2018) describe. Another one can be Facebook's "Why am I seeing this ad?", a tool for people to better understand the reasons they were being shown a particular ad in the platform (META, 2022). This is what she would call *explainability.*

It is important to note that Rudin (2019, p. 4) traces a harsh criticism on the notion of *explanations*. According to her, "many of the methods that claim to produce *explanations* instead compute useful *summary statistics of predictions* made by the original model". In this sense, the information provided by these systems would not consist of an "attempt to mimic the calculations made by the original model", but in reality a display of trends in how predictions are related to the features. For this reason, they could be often not reliable, sometimes misleading.

Rudin's differentiation is an interesting way to assess computational instances of different transparency techniques and methods. However, as the purposes of this study relate more to a study of ML transparency from the perspective of regulation and policymaking, and does not enter deeply into the details of how computational operations to put into practice such processes and outcomes take place, we will, just like Miller (2018) and Hamon et al. (2022), use both terms here interchangeably. However, in a difference sense: interpretability and explainability will refer in this work to the *techniques and methods for creating and enhancing the level of transparency of ML systems*. As mentioned above, they are, thus species of the larger gender *transparency*.

For the sake of completeness, it is important to notice at this point that some stakeholders have also been talking about a right to explainability that would exist in both

the European Union's General Data Protection Regulation - GDPR (EDWARDS; VEALE, 2017) and the Brazilian General Data Protection Law - LGPD (WIMMER, DONEDA, 2022).

*Sidenote: ML components*

One further conceptual note should be made with regard to machine learning systems' components. As the concept of algorithm does not encompass the data applied for feeding the system, but only the set of commands that leads to its output, we will refer to the term *model* to encompass both the data and the set of instructions (algorithm). As we will further discuss, assessing how data is processed by a system is especially important for complying with data protection regimes such as the LGPD in Brazil and the GDPR in Europe. Finally, when talking about *system*, we refer to the ensemble of data, algorithms, outputs as well as the social, environmental, economic and political dynamics involving these systems.

### 3.2. Promoting transparency: a brief set of examples of interpretability/explainability methods

Methods for promoting interpretability/explainability may be classified according to three criteria: moment, scope and degree. The first we assess refers to the **moment** when the method is applicable with regard to the building of an ML model: before (pre-model), during (in-model) or after (post-model) the issuing of an output of the system.

*Moment-based methods*

Pre-model explanation techniques are developed without thorough consideration of the model itself and are thus independent of it, applying only to the data which is to be used to feed the system. They relate to techniques for data visualisation (CARVA-

LHO, PEREIRA, CARDOSO 2019), such as t-SNE, for example, which allows for high-dimensional data visualisation by giving "each datapoint a location in a two or three-dimensional map" (MAATEN & HINTON, 2008), or Principle Component Analysis. Under the perspective of data protection regimes, such an approach is important so as to assess which personal data is being processed by the system. However, it does not allow by itself to understand how such data is being used.

In-model approaches, on the other hand, relate to models which are inherently interpretable for having been embedded with tools for explaining their functionalities from their very development. They aim to answer the question of "how does the model work" (GILPIN, BAU, YAN, 2019) and, consequently, how they process training data.

Post-model techniques, on the other hand, concern the improvement of a system's explainability after it has already been built (CARVALHO; PEREIRA; CARDOSO, 2019). Most post-model techniques are also post-hoc, meaning that the model is explained after it has already been trained, and aim to answer the question of "what else can the model tell us". According to Lipton, "one advantage of this concept of interpretability is that we can interpret opaque models after-the-fact, without sacrificing predictive performance" (LIPTON, 2016).

*Scope-based methods*

Techniques for promoting intelligibility can also be categorised according to their **scope**, which "refers to the portion of the prediction process they aim to explain" (CARVALHO, D.V.; PEREIRA, E.M.; CARDOSO, 2019). They can provide either algorithmic transparency or global and local explanations.

Algorithmic transparency (not ML transparency as a whole) allows one to comprehend how the algorithm learns from data and what kind of relations it can extract from such an operation. In this sense, algorithm transparency's goal is to learn how the algorithm works, and not individual predictions. It does not require knowledge about the data or the learned model, but strictly about the algorithm itself, which is,

the set of instructions which allows the system to perform a specific task. Hence, it is a way to answer the question of "how does the trained model make predictions?" (MOLNAR, 2019)

Differently, global interpretability is applied when the agent's goal is to describe the behaviour of the entire model, which includes an understanding of the data and the algorithm itself (ARYA et al, 2019). This explanation may be holistic or modular. Global holistic model interpretability aims to explain the entire model at once and understand how it makes predictions, which requires knowledge of the algorithm and the training data. A model can only be holistic if it is simple enough, since any model that has more than five parameters or weights is unlikely to fit into the short-term memory of the average human (COWAN, 2010), and is thus very difficult to be achieved (MOLNAR, 2019).

Global model interpretability on the modular level is more practicable to be achieved. They do not aim to explain every single feature of a machine learning model, but instead to explain the model by separating specific features used in decision-making processes and trying to understand how they worked. The question to be answered by this method is "how do parts of the model affect predictions?" (MOLNAR, 2019)

Finally, local model interpretability aims to describe single predictions, and may be reached through explaining (1) a single prediction, which can be done by zooming in on a single instance, examining what did the model predict after processing a specific input and explain why; or explaining (2) a group of predictions, by selecting a group of instances and understanding how does the model make specific predictions for this group (ARYA et al, 2019).

Whether we need global or local explanations depends on the information an individual needs in order to reach a specific goal. For instance, in order to understand what role a recommendation system plays in a social network to profile users and display personalised content, a regulator would probably make better use of global model explanations. The regulator's goal would be mostly to comprehend how the system works in general, so as to develop more effective regulation to guide platforms in the development of algorithms that better identify misinformation online and rapidly re-

spond to it by, for example, reducing the reach of a specific content. Algorithm transparency could also be of good use in this context, since understanding the chain of commands may show itself useful for identifying eventual biases in its metrics (BOZDAG, 2013).

On the other hand, a user who does not want to receive specific advertisement in a search engine would probably make better use of a system which explains how it targeted such content to the user, based on which inputs and the weight of each of them in that particular recommendation. In this sense, a local explanation would possibly be a better fit.

A local explanation might also prove itself useful for assessing whether a credit-scoring system has been biased or not when rating a specific person. Understanding what data it used and how the system performed to take this particular decision might be points for both regulators and users scrutinise.

*Degree-based methods*

Another taxonomy for explanation relates to the **degree** to which a user may interact with the model, and they might be static or interactive. A static explanation does not change in response to user's feedback. On the other hand, an interactive one allows users to dialogue with the model to request further information on a decision made, such as by drilling down or asking for different types of explanations (e.g., through dialogue) until they are satisfied (ARYA et al, 2019).

Based on the above, we can reach the following table with transparency techniques depending on their moment, scope and degree:

| Moment | Scope | Degree |
|---|---|---|
| Pre-model | Algorithmic | Static |
| In-model | Global model | Interactive |
| Post Model | Local model | - |

Table 1 - Criteria for interpretability/explainability methods

Having observed how explanations are usually provided, it is now appropriate to understand what are the arguments brought by scholars from different fields when discussing the effectiveness of ML transparency in addressing the risks and impacts of these systems. We aim, hence, to answer the following questions: is ML transparency indeed effective? If so, for what, and to what degree?

### 3.3. The Ups and Downs of ML Transparency

*The Ups*

It has already been some decades since those involved in the developing of computational systems have noted that models should not only be accountable for their operations but that they should also allow for users to understand how they reached their outputs (VEALE, 2019). In 1976, Edward Shortliffe (1976) signalled that making these machines issue explanations about how they reached their outputs could be an important means for them to gain human trust. He thus described how implementing *explanation facilities* in these systems would be an important step towards this goal.

Years later, in 1997, Dourish (1997, p. 11) highlighted the potential for making systems explain themselves through what he called *accounts*. They would be "causally-connected representations of system action which systems offer as explications of their own activity". At that time, he drew attention to the fact that systems' explanations are not neutral, but inevitably select and hide specific information about their functioning in order to be more meaningful for their understanding by a human being.

Differently from most of the views presented today, the explanation facilities described by the aforementioned work of Shortliffe (1976) were mostly focused on making the systems understandable by those who were administering the decisions, not on the ones being impacted by it. According to Veale (2019, p. 56-7), this approach was justified by two main reasons.

The first is that, at that time, Shortliffe's focus was on the development of so-called "expert systems", AI systems not based on machine learning techniques which aimed at mimicking the work of experts from different fields such as in medical diagnosis, credit card fraud detection or chess playing (OECD, 2022). Due to the difficulty in moving these expert systems from research to deployment in the 1970s, the acceptability of the specialists that would use them was seen as more important for their development than that of those subject to the systems' outputs - in the clinician system case, the patients. The second reason was because these machines were designed not to create a fully automated process, but to turn lay users into specialists or to enhance the speciality of experts.

With the growth in ML complexity and adoption, this scenario has changed. ML started to be more applied in the most diverse fields, by people with different degrees of understandability of its functioning. They went gradually far beyond expert systems. As their effects extended beyond the walls of university labs and towards the whole society, scholars and activists have raised ever increasing calls for rendering them more transparent towards no longer only experts, but especially for the subjects who were being affected by them. Individuals should, according to this view - or "ideal", as Veale (2019, p. 58) puts it -, have effective control of algorithmic decision-making. At the same time, transparency became ever more difficult to be rendered due to the increased quantities of data used for the training of these systems and for their growing complexity, increasing the opacity in ML and creating the so-called black boxes (ASGHARI et al., 2021).

Frank Pasquale's book The Black Box Society is a frequently cited work when scholars make reference to ML opacity. It describes how massive personal data processing operations by algorithmic systems and the outputs that they issue have been made under a frequently wilful veil of secrecy that hides, within metaphorical, digital black boxes, the values and prerogatives enacted by the encoded rules. Pasquale assumes that ML systems are an example of a digital one-way mirror, through which ML operators can see through individuals' data and extract meaningful knowledge about them while persons have no clue on how such extraction is made in practice and how do they impact their lives. As the deployment of these systems have been exerting a thor-

ough degree of authority over populations, *opening these blackboxes* to individuals and/or regulators would be a tool to making society more transparent, and is crucial to understand how fair these machines are (PASQUALE, 2015).

Beyond Pasquale, other scholars have expressed how transparency could be seen as a path towards enhancing ML accountability. Doshi-Velez and Kortz (2017), for instance, see explanations as a way to prevent errors and increase trust by exposing the logics behind a specific output. It would also be a useful mean to determine whether certain criteria - such as personal information - were used appropriately or inappropriately by the system. In this sense, they consider "when and what kind of explanation might be required of AI systems" (DOSHI-VELEZ & KORTZ, p. 2) as the main questions the scientific and policy community should answer.

Similarly, Rader, Cotter and Choo (2018) argue that transparency can (1) create awareness that interactions with a system are mediated by an algorithm; (2) help users learn more about how the system works in order to evaluate whether its outputs are reasonable or not; and, ideally, (3) enable users to identify biases and empower them to question and critique a system.

Edwards and Veale (2017) defend that there might be cases in which information regarding the functioning of a system can be essential to mount a challenge against the deployer of a problematic ML system in court or to a regulatory body. Disclosing such information could thus be a tool for individuals, civil society organisations, researchers and regulators to combat harmful or anti-competitive practices involving ML by both private and public agents. As we will see below, however, they are quite skeptical of how effective such transparency might indeed be.

We can resume the views expressed by these scholars mostly to the idea that transparency has been mostly defended as a means to both increase trust in systems and also allow for them to be accountable. As Marda (2018) puts it, calls for transparency have been mostly focusing on the intelligibility or scrutability of ML systems in order to understand and study their behaviour. She adds that transparency should ultimately allow for a reconfiguration of power structures that push developers farther from a

position in which they hold all the cards for accountability and thus "choose to be accountable" (MARDA, 2018, p. 6).

*The downs*

Despite the arguments presented above referring to a potential effectiveness of transparency as a tool to promote accountability and increase, a large part of the scholarship seems skeptical on how much effective can ML transparency be to increase trust and enhancing accountability (HAMON et al., 2021).

On the one hand, authors such as Linardatos, Papastefanopoulos and Kotsiantis (2020), for instance, argue that "there is clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions", especially when one aims to understand the functioning of deep learning models. On the other, there are those who have been putting into question this assumption, such as Rudin (2019), which draws attention to the fact that interpretable complex models are not only possible but also necessary, especially when they are applied to high-stake decisions in fields such as healthcare.

At the same time, HAMON et al. (2022) identify that there is a clear trend in machine learning development that tends constantly to higher complexity, caused by three main factors: (i) an increase in data complexity, due to the feeding of ML systems by data sets that are getting gradually more heterogeneous and high-dimensional; (ii) the use of an ever growing number of parameters in these systems; and a (iii) growing sophistication of algorithms and techniques used for the development of these machines. As a result, explanations about these systems' functioning would also tend to become ever more complex, something which raises doubts as to how understandable they might actually be.

Beyond these more technical questions, related to whether a system can technically be rendered transparent or not, scholars struggle to reach consensus especially when analysing whether the relationship between transparency and accountability is as obvious as it is commonly described. Despite a certain relationship between these con-

cepts that traces back to calls for the transparency of state activities in order for society to exert control over the acts of rulers (ANANNY & CRAWFORD, 2016), they are not synonyms.

Accountability refers to the relationship between an agent and a forum in which the former has "an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences" (BOVENS, 2006, p. 9).

Transparency, on the other hand, is "only the beginning of this process" (EDWARDS; VEALE, 2017, p. 41), or even merely one out of many other means for achieving effective accountability, specially concerning ML systems (DOSHI-VELEZ & KORTZ, 2017). The ideal of transparency has thus its limitations, and has a direct correlation with contextual, relational ways of disclosing information.

Ananny and Crawford (2016) have argued that the focus of the computer science (and also those in the policy debate) community on transparency has been excessively driven towards an ideal that merely *opening* these systems and seeing inside their inner functioning would be enough to render them accountable. For the authors, these groups often see transparency as a way to bring insight and governance in the deployment of ML systems, by assuming that the accountability of objective computational technologies like algorithms could be enacted by the merely looking at their technicalities, such as source code or the databases used for their training.

As we saw earlier, the view of opening black boxes has been the centre of a strong advocacy effort in scholarship by authors such as Pasquale (2015), but not without criticism, such as the one we mentioned in Chapter 2 made by Bucher (2018) regarding the flaws of referring to ML systems using the black box metaphor.

What is at the core of the critic of Ananny and Crawford (2016), however, when they question whether is it indeed possible to open these black boxes and, if so, whether this would be sufficient, is the implicit assumption that merely seeing a phenomenon is enough to create opportunities and obligations to make it accountable and thus change it. Not necessarily understanding a system allows one to effectively change the way it works or even make it not be used at all. For this reason, the authors argue

that scholarship should shift towards an idea of, rather than looking *inside* such systems, looking *across* them. This would mean to see ML technologies as "sociotechnical systems that do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans" (ANNANY; CRAWFORD, 2016, p. 974).

The authors, in this sense, defend that transparency is far from allowing for an effective form of control. Seeing, having access to a given object, is not by itself enough for granting the necessary knowledge to render it accountable, and this relates not only to algorithms, but to many other social systems. Despite the fact that the access to a system's inner workings can indeed provide insight and spur further investigation, Ananny and Crawford (2016, p. 978) argue that "significance and power is most revealed by understanding both its viewable, external connections to its environments and its internal, self-regulating workings". They thus trace a series of limitations of the transparency ideal that put into question the potential benefits that we traced earlier in this section. We summarise some of them in the following paragraphs.

First, they argue that transparency is useless if there is no effective system for processing a disclosed set of information that denounces the vicious of an agent in a way that produces change. In these terms, when unethical or unlawful actors are not vulnerable to the public exposure, the transparency mechanism can, counterintuitively, lead to more cynicism and, consequently, to greater public distrust, as there are no effective tools to render bad actors accountable (ANNANY, CRAWFORD, 2016,).

This lack of trust can also dwell among the ones being watched, when they do not trust the ethics and intentions of those who might have access to the information (ANNANY, CRAWFORD, 2016, p. 978, 980). This argument is strong among the corporate sector, specially when it applies to trade secrecy. Private actors argue that opening up to public scrutiny would allow competitors "to free-ride on innovator-based technology and reduce the latter's competitive edge", creating thus a chilling effect on innovation. On the other hand, a larger opening of companies' information would allow for bad actors to compromise their operation by manipulating or exploiting vulnerabilities (ALÌ; YU, 2021, pp. 6-7).

Transparency can also induce *harm* when it is used without an assessment of why and how a system should be disclosed, opening the path to violate privacy rights by opening carelessly information on persons or specific groups (ANNANY, CRAWFORD, 2016, pp. 978-9).

Transparency can also end up *occluding* when the amount of information is so great that one cannot assess it effectively. That can happen either intentionally, by what Stohl et al. (2016) call *strategic opacity*, or unintentionally, through an *inadvertent opacity*. (ANNANY, CRAWFORD, 2016, p. 979)

Another limitation for the transparency ideal lies on the placing an expressive *burden* on individuals to interpret information, invoking "neoliberal models of agency" that require responsibility from the ones in the weaker side of society to control wrongdoings usually in a situation of severe information asymmetry (ANNANY, CRAWFORD, 2016, p. 979).

Transparency can also "*privilege seeing over understanding*" (ANNANY, CRAWFORD, 2016, p. 980), when systems are only made visible but are not debated or challenged by observers who have enough knowledge and are in a position to do act upon them.

Ananny and Crawford go on with their argument by also affirming that transparency have both technical and temporal limitations. In the case of ML systems, a frequent example of technical limitations involves deep learning systems such as the ones used in image recognition programs. This sort of system is usually so complex that in most cases even their developers cannot precisely say what caused their mistakes. Temporal limitations, on the other hand, are expressed in the idea that "different moments in time may require or produce different kinds of system accountability" (ANNANY, CRAWFORD, 2016, p. 982).

When applied to ML systems, this situation can be identified considering how systems change over time, both through frequently gathering new information from training data and interacting with other systems. An example of this can be seen in content recommendation systems in social media platforms, where multiple, different algorithms are always in an interplay with one another (BUCHER, 2018).

The limitations expressed not only by Ananny and Crawford (2016) and the other scholars presented above lead to an understanding that they are not just code and data, but an "*assemblage* of human and non-human actors" surrounded by particular power dynamics, preconceptions and economic and political interests. In this sense, the authors call for a consideration of the limitations of transparency as a starting point for identifying what makes accountability effective in a specific environment and what are its own limitations, going beyond transparency.

*A middle way: towards qualified transparency*

We may interpret the argument of the aforementioned authors as a certain form of skepticism, or even disillusionment, towards the idea of machine learning transparency as a panacea, as something enough per se to hold those responsible for ML systems accountable.

This notion makes sense when talking about the disclosure of information regarding what we may call both the internal (the inner-operation, or what we also have called "technicalities") and external (the market, policy choices, as well as the humans involved in the system's development and deployment) environments of ML. Transparency should always be thought as dependent on the tools and knowledge that will allow it to drive a subject towards effective accountability on a case by case basis. The red lights turned on by the authors we mentioned in the last pages are precise in drawing attention to the limitations of the transparency ideal, or even, going back to Veale (2019), the ideal of effective control over the technology that transparency by itself would enable.

However, one should not disregard how ML transparency has already allowed for important unveilings of ML's flaws, many of them by independent investigators, as we referred earlier in this chapter.

Case studies have already revealed discriminatory biases in systems used in health care management programs (OBERMEYER; MULLAINTHAN, 2019), in self-driving cars that had issues with detecting pedestrians with darker skin colour (WILSON

et at., 2019), and the famous ProPublica case whereby researchers and civil society advocates discovered that the COMPAS algorithm, used to predict the likelihood of recidivism for on-parole probation, discriminated against African-American convicts (LARSON *et al*, 2016). In most of these cases, researchers could not have achieved these discoveries without having access to important information regarding the systems in use, and access to the training data sets was of particular importance in these examples (ALÌ; YU, 2021).

This goes hand in hand with the argument that, sometimes, we should be "less concerned with providing individual rights on demand to data subjects" and more concerned with "empowering agencies, such as NGOs, regulators, or civil society scrutiny organisations, to review the accuracy, lack of bias and integrity of a ML system in the round and not simply challenge ML decisions on behalf of individuals" (EDWARDS; VEALE 2017, p. 23).

Transparency, thus, seems to depend on many different conditions in order to be truly effective in safeguarding the rights of persons and communities affected by them. Issues related to trade secrecy, privacy, system's characteristics and the effectiveness of the enforcement of the rules by authorities are some of the aspects to be considered when conceiving ways for promoting ML transparency that is useful.

Nevertheless, this does not affect the importance of transparency as one among multiple tools necessary to address accountability. As affirmed by Asghari et al. (2021), "[n]o one should be subjected to norms and institutions which cannot be justified towards them, based on reasons which they cannot question".

That is why Frank Pasquale (2015) affirms that what we need is "qualified transparency". Most of the times, it is not having access to a system's code that will help us solve an issue relating to the functioning or the surroundings (environmental costs, predatory business practices) of an ML system, but instead have access to "limiting revelations in order to respect all the interests involved in a given piece of information" (PASQUALE, 2015, p. 142). After all, source codes will be probably useless to learn what is the carbon footprint of an ML system; or to understand to what measure

a social media algorithm amplifies the reach of a post containing hate speech. Instead, this will be provided by information delivered by a developer or deployer through record keeping or other means. And preferably through their own will, so that we do not have to always depend on whistleblowers putting their careers at risk to reveal to society what are the harms of ML systems used by organisations.

In sum, although being crucial, transparency has its limitations and a direct correlation with contextual, relational ways of disclosing information. Further, it should be seen as valid only to the extent to which it allows for effective action and reflexivity about how ML helps perpetuate exclusion, including through racism, colonialism and misogyny (D'IGNAZIO, KLEIN, 2020; SILVA, 2022).

With that in mind, the next section aims to understand what are the questions that should be made in order to understand what sort of transparency one has in mind and the nature of the information one needs when assessing the lawfulness of an ML system and the actions of those developing and deploying it.

### 3.4. Questions to be answered

As one may conclude from the previous pages, the effectiveness of transparency mechanisms will vary according to multiple variables, and are thus context-dependent. According to Hamon et al (2022), explanations about ML systems are contextual, as "each form of explanation (ex-ante, ex-post, expert-oriented or subject-oriented, more or less granular) strongly depends on the context of the problem and on the capacity of the data subject to interpret the results".

As such, authors such as Hamon et al. (2022) propose that a series of interrogations should be carried out in order to identify what main features a person should take into consideration when demanding information about an ML system. According to them, the design of explanations regarding the functioning of ML systems would depend on four main factors: the *moment* of the disclosure of the explanation, if either before (*ex ante*) or after (*ex post*) the issue of an output; the *audience* of the explanation, i.e., the explanation will change if it is made for an expert or for a data subject which may be

affected by the system's output; the *layer of granularity,* such as if the explanation should be made in a similar way to anyone assessing it or if it should be group- or individual-based; and the *level of the risks* of the automated decision regarding fundamental rights and freedoms.

Doshi-Velez and Kim (2018) recommend a complementary set of questions for assessing the usefulness of the information provided about a system, shedding light into four different aspects to be taken into consideration when assessing the transparency level of a system.

Their first question is: *does one need to understand the whole system or a specific decision*? In the former case, the global interpretability we saw above would be necessary, while on the latter local explanations would seem enough. With regard to content recommendation systems, for instance, such as the ones used in social media or search engines to select and hierarchise information for users different goals would make each of them useful: if one needs to understand how an algorithm ranks content according to users' engagement with them in a more generalised way, a global explanation would be required. However, if a user intends to know why she had access to a specific piece of misinformation about a miraculous cure for COVID-19, she would probably find it more interesting to know what personal data did the system take as a parameter to reach the output (which is, in this case, the display of the misinformative content).

The second feature would be the characterisation of incompleteness, which aims to answer *what part of the formulation is incomplete, and to what extent*. This relates to which information explanation is required, such as knowing about the dataset used in a facial recognition application which seems to have made a biased identification, or in case one needs to know about the image captured by the system's camera. Each case will require a different form of explanation (DOSHI-VELEZ; KIM, 2018).

The third relate to time constraints: *how long can the user spend trying to understand the explanation*? Back to our content recommendation system, if a user wants to know why a misinformative content was displayed, the person may want to spend no more than a few minutes minutes understanding the explanation. Differently, a researcher

looking for biased datasets in facial recognition applications would probably be willing to spend maybe hours analysing data to find a response (DOSHI-VELEZ; KIM, 2018).

The final aspect pinpointed by Doshi-Velez and Kim (2018) relates to the nature of the user expertise: how experienced is the user for understanding this sort of explanation? For a doctor trying to understand why an ML model classified a tumour based on the image of a patient would probably require a level of sophistication from the explanation method different from the social network user who does not need a complex understanding of the nature of the profiling being carried out about her. The type of language required would also be different in these examples: while the doctor would find it crucial to receive the information in technical, medical wording, plain language would probably be a better fit for the social network user.

The aspects herein highlighted may prove themselves useful for regulators when assessing whether and how a specific piece of information should be provided so as to increase the intelligibility of a machine learning system. Based on the works of Hamon et al. (2022) and Doshi-Velez & Kim (2018), and also on the ideas exposed throughout this study, we would summarise these variables, all of them necessary to shape what information about the ML system is indeed helpful and how should it be delivered, in six main questions.

*Tracing our own questions*

A first question to be answered by an individual or group that aims to understand an ML system relates to the *subjects* to whom the information is necessary. These persons and groups usually have different goals, origins, intellectual backgrounds and interests, to say the least. As we have seen, Shortliffe (1976) was worried about having his systems understood by physicians; Pasquale (2015) highlighted that regulators, law enforcement agencies and individuals should have access to information regarding ML systems that have relevant impacts on society in order to promote accountability; Alì and Yu (2021) drew attention to the importance of computer scientists and civil society organisations having access to these systems in order to identify potential

pitfalls; the European Union's Digital Services Act (EUROPEAN PARLIAMENT, 2022) includes auditors as potential parties with an access to social media algorithms' source codes.

Looking at these different groups of information recipients, one can first realise that their knowledges and interests vary. The physician might be an expert in oncology, but may not know much about how a computer vision system. Individuals may have different education and digital literacy levels, pertain to varying age groups or social classes, may come from backgrounds as diverse as Brasília or Beijing, be part of different ethnicities, and so on. Regulators may be from fields as different as banking and disease control, and be interested in information that range from what datasets were used to train a system, who participated in its development or deployment, what are the weights of each information when a system releases an output and the economic and political choices that led to development of an ML model. And the list goes on.

In this sense, knowing *who* is the recipient of the information is an important part of the assessment not only of what is meaningful to know about a system, but also *how* should this information be shared with the subject, with what kind of jargon and in what circumstances, for example.

On top of that, it is also fundamental for an assessment of how should information on an ML system be delivered that one knows what are the *purposes* that a subject has when requesting access to such data. A lawyer might be interested in mounting a legal challenge. An individual may want to understand why he or she has been approached by the police after the deployment of a face recognition system; a banking sector regulator might be interested in how does a bank's credit scoring system works; a regulator how a search engine's recommender system algorithm works to assess whether and how it may be impacting how people access information during an electoral campaign.

This leads us to the question of *what kind of information*/data is necessary for achieving these intended purposes vary a lot according to each specific situation. Access to the database would be enough? Or one would needs to understand how a specific de-

cision was made by the system? In this sense, a counterfactual explanation would be enough?

The data in which someone is interested may also go beyond the functioning of the system itself, and range from how was the bidding process through which a state in Brazil or the United States bought a face recognition system; the level of bias in a credit scoring system dataset; or how a social media allows its users to personalise its feed and what data is being used for what purposes. And of course, this will also vary according to the *kind of system being used* and the *risks* it poses, our two last questions, as the information provided will depend, firstly, on the particularities of each system and, secondly, on how they may affect individuals and communities's rights and interests.

Systems which may not have any impact on individuals or the environment may not necessarily be required to provide complex explanations. However, as it might be frequently the case that we will sometimes not immediately identify the impacts of systems right from their development or placement in the market, we will sometimes require further information about systems that have apparently low or no risk. Moreover, risks are usually defined in not neutral ways that are not necessarily future-proof and thus change over time (BALDWIN, CAVE, 2021). For this reason, risk should not be seen as an ultimate determinant to require further transparency, but instead as a supporter in understanding what to prioritise in enforcement, in a way that does not leave supposedly low risk applications unwatched. In this sense, on the one hand, the regulator should always have flexibility in defining how it assesses risk and, on the other, those creating and deploying ML systems should be attentive and ready to provide information about them when so requested.

With that in mind, a further summarisation of these questions may be expressed as follows:

1. ***Who*** is the recipient of the information? A regulator? A lawyer? An individual or community affected by the system? A civil society organisation? An expert on the specific field where a system is being deployed? A computer engineer/scientist?

2. ***How*** should the information be provided? Is it necessary for the system to be audited or is an explanation enough? In case of the latter, can the information be transmitted using expert vocabulary or in a way that a layperson should understand?

3. ***For what*** purposes? Is it to mount a legal challenge? To fix a bug? To assess the fairness or bias of an output? To assess the reasons for a system's output that claims that a patient has cancer?

4. ***What kind of information***/data is necessary for achieving the intended purpose? Is it enough to assess the training dataset, general information about the operation of the system, or, going beyond the system itself, its energy expenditure or the economic and political choices that led to the application's development? It is important to know how the data was collected, such as directly by humans, automated sensors or both? Further, does one need information of the system as a whole or how it achieved a specific decision, such as why it refused credit to an individual?

5. ***What kind of ML system*** is being assessed? Is it used for e.g. face recognition, social credit calculation or to personalise content online?

6. ***How risky*** is the ML system for individuals, groups or society as a whole?

The questions posed are a first step for the assessment of the necessary information one need to access regarding an ML system and the contexts in which they are built. They are a path towards understanding not only the technical details of a system, but also the environments in which they are developed and used, as well as the power dynamics, the economic and political interests, and the biases of the ones involved in their creation, selling and deployment. As Silva (2022) argues, going beyond code is fundamental to allow for the combat against algorithmic racism, and we should add here any kind of technological oppression.

Being able to make such an assessment is, in this sense, important both for the ones who are involved in ML systems' development and deployment, as well as for regulators to assess how effective is a model in providing meaningful information about its functioning.

A useful way to assess the effectiveness of the information provided about a system is by assessing, as proposed by Carvalho, Pereira and Cardoso (2019) whether the explanations achieve three goals.

The first one is accuracy, which relates to having a connection between the prediction made by the machine learning model and the explanation provided by the explanation method. It is important because it is possible to have an understandable hypothesis which has no connection to the data.

The second is understandability, referring to how easy an explanation is to be understood by the user. Especially with regard to recommender systems explainability, such as in a social network or a search engine, for instance, most users are not skilled in computation, the explanation of a decision should be easy to understand in order to be minimally helpful.

Thirdly, an explainable method should be efficient, which means that it reflects the time necessary for a user to fully grasp the explanation. It refers thus to how understandable it is in a finite and preferably short period of time.

These three goals can be an effective way to assess how useful is the information provided, and thus help understand whether it has proven effective or not for the stakeholder's goals. In addition, as we will see in the next chapter, Kaminski (2019) advocates for the provision of information that are understandable, meaningful and actionable. By actionable we may comprehend a more normative stand, in the sense that individuals are able to take action in reclaiming their rights if, through the information provided regarding the ML system, they identify the violation of rights.

In this chapter, we paved the way to first understand what key concepts used in the field of ML transparency mean, such as transparency itself, interpretability and explainability, in order to clarify what each of them imply. Afterwards, we analysed different perspectives on whether and to what extent ML transparency is effective in addressing ML's impacts and enhance accountability. We finally reached the conclusion that transparency, although being the means to different ends and not enough per se to solve ML's problems, is an important tool to be taken into consideration by society to combat the problems raised by ML. Our analysis later analysed different ways in

which transparency and, more specifically, interpretability and explainability are usually put into practice, up to the point at which we reached what are the necessary questions to be asked by a stakeholder when demanding transparency and assessing its effectiveness.

As a next step, it is important to acknowledge that ML calls for transparency can involve to a great degree the protection of personal data and the legal regimes that relate to this right. As we have shown above, transparency has been a demand that responds to the impacts that these systems have on individuals, groups and the environment, specially when they issue outputs that to some degree are the result of personal data processing.

With that in mind, as we look at the problem of ML opacity from a legal, regulatory perspective, it is inevitable to assess how data protection regimes deal with the issue of ML transparency, and we focus on the ones enacted in Brazil and the European Union. Although not having ML as the central object of their provisions, the fact that by the time of writing the AI legislations are still under discussion in these territories, and that data protection regimes in these territories affect the processing of data by ML systems, these laws are maybe the most suitable legal regimes already enacted and into force in Brazil and Europe to assess ML transparency. Adds upon this the fact that they have specific provisions related to the provision of information regarding automated decision-making systems, which, as we will further see, are an umbrella that includes machine-learning. In this sense, it is inevitable to take them into consideration if we want to understand the enforcement of transparency. That is what we will do in the next chapter.

## 4. MACHINE LEARNING TRANSPARENCY AND DATA PROTECTION REGIMES

In the pages above, we saw that making machine learning systems transparent and understandable can be an important tool both for addressing the accountability of agents involved in their development and deployment and understanding how they are changing us and the ecosystems in which we as humans dwell. Although not sufficient per se for achieving effective accountability, transparency is an important step towards understanding whether and how an ML system can impact people and the environment.

When we assessed how machine learning technologies work in Chapter 1, we saw that their development depends mostly on the processing of data. Based on the input they receive, they are programmed to look for patterns on the information and, following a well defined or not set of rules, issue outputs that can be recommendations, classifications, summarisations, regressions or whatever one may aim to have.

Frequently, ML systems process personal data as either input or output. Systems used in credit scoring, face recognition and content recommendation in social media are some examples of technologies that process information related to or which may relate to persons.

As such, when researching ML governance and regulation, it is not easy to avoid looking into this theme from a data protection perspective, especially considering that, at least to the point of writing this work, general artificial intelligence legislations have not been approved in the jurisdictions that we aim to look at, namely Brazil and the European Union (EU). In this sense, one may find particularly noteworthy the remarks made by the United Nations' Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression that data protection laws can be one way to regulate AI (and thus machine learning) with existing norms, in particular if made more flexible (UNITED NATIONS, 2018, p. 16).

With that in mind, in this session we will explore how data protection legal regimes in both legal frameworks address machine learning transparency. We will take them as a

reference for how law may be used as a tool to push those involved in an ML system life cycle to promote further transparency regarding these technologies. These laws are Brazil's *Lei Geral de Proteção de Dados* (General Data Protection Law, LGPD, Law no. 13,709/2018) and European Union's General Data Protection Regulation (GDPR, Regulation no. 679/2016).

One may reasonably note that other legal instruments in these jurisdictions touch upon issues related to machine learning. In Brazil, an example is the *Conselho Nacional de Justiça* (National Council of Justice, CNJ)'s Resolution no. 332/2020, which establishes rules for the use of artificial intelligence systems by judicial bodies (BRASIL, 2020a). Moreover, proposals still not approved, such as the *Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet* (Brazilian Law on Freedom, Responsibility and Transparency on the Internet, Proposal no.2,630/2020) (BRASIL, 2020b), and the *Marco Legal para a Inteligência Artificial* (Legal Framework for Artificial Intelligence, Proposal no. 21/2020) (BRASIL, 2020c)[2] are other instruments that, when approved, will probably have large impacts on the ruling of ML systems.

At the EU, the Database Directive (Directive 96/9/EC) (EUROPEAN UNION, 1996) may serve as a legal means to assign property rights over datasets key to training machine learning systems, according to Veale (2019). The recently voted by the European Parliament Digital Services Act (EUROPEAN PARLIAMENT, 2022) also has specific provisions related to the governance of algorithmic systems (including machine learning) applied in social media and other digital platforms for recommending content. The AI Act (EUROPEAN UNION, 2020), proposal yet under debate at the European Parliament, has the regulation of these systems at its core, and, once ap-

---

[2] After intense criticism from civil society, the Rapporteur for the Senate of the *Marco Legal para a Inteligência Artificial*, Senator Eduardo Gomes, commissioned a Commission of Jurists to prepare a new draft for the *Marco Legal para a Inteligência Artificial* on the 30th of March, 2022. The bill's scope is expected to considerably change after the presentation of the new text. See, e.g., LEMOS, Alessandra et al. **Nota Técnica PL n. 21/2020**: sobre o marco legal do desenvolvimento e uso da inteligência artificial no brasil. LAPIN, Brasília, v. 1, n. 1, p. 1-49, nov. 2021. Available at: https://lapin.org.br/2021/11/09/nota-tecnica-atualizada-discute-o-pl-21-a-2020-do-marco-legal-de-ia/. Last access: 12 jul. 2022; DE PEREIRA, José Renato Laranjeira de; MORAES, Thiago Guimarães. Promoting irresponsible AI: lessons from a Brazilian bill. **Heinrich Böll Stiftung,** Brussels, 14 Feb. 2022. https://eu.boell.org/en/2022/02/14/promoting-irresponsible-ai-lessons-brazilian-bill. Last access: 12 Jul. 2022.

proved, will also significantly impact their development and deployment.

Nevertheless, having briefly spoken about these legal instruments, the main reasons for still choosing to ground this work on a comparative analysis between data protection laws are:

i. Their hierarchical superiority compared to other already approved legal instruments that may have a regulatory power over ML in these jurisdictions. That when considering the Brazilian case, where the aforementioned CNJ's Resolution no. 332/2020, although already into force, has an inferior legal status and thus more limited range of application (applies only to the use of ML by courts) when compared to the LGPD;

ii. The general, broad scope of the data protection laws, since they may apply to any personal data processing carried out by entities from both private and public sectors, differently to Brazilian Proposal no. 2,630/2020 or European Union's Digital Services Act, which have specific addressees to their provisions and, at least in principle, all of them from the private sector;

iii. The fact that they are already into force, in opposition to Brazil's Proposal no. 21/2020 and EU's AI Act, which are yet under debate by lawmakers in both jurisdictions;

iv. Some of the most important and worrisome machine learning systems involve the processing of personal data at some point in their development and deployment, including during their training or when applying their results to individual situations (VEALE; 2019), and are thus, in these cases, subject to the obligations derived from data protections regimes in both Brazil and the European Union.

Based on the above, we will carry out in this chapter a comparative analysis of data protection regimes in both jurisdictions. Our focus will be in understanding how legal provisions in the LGPD and the GDPR address, directly or not, the provision of information regarding machine learning systems.

With that in mind, we will split this chapter in three parts. The first will consist of an exploration of what data protection means and how it historically came into being and became a right in the European Union and in Brazil. We will analyse how the debate on the right to privacy in the US evolved until the conception of the right to data protection and its further development in Europe and then in Brazil. The questions we aim to answer in this section are thus "what is data protection? How did it come to being? How did it legally evolve?"

The second section starts diving into the LGPD and the GDPR content. It assesses what is the scope of these legal documents, what are their key concepts and how do they specifically apply to the processing of personal data by machine learning technologies. We thus expect to answer "what are the LGPD and the GDPR about? What are their main principles, rights and obligations? How do they define personal data? How do they apply to personal data processing activities carried out with machine learning systems?"

Finally, the third section aims to interpret the rules established by these laws regarding the provision of information related to personal data processing activities carried out by machine learning systems. Its goal is to understand what rules regarding ML transparency may one extract from the LGPD and the GDPR. Our final purpose is to interpret the provisions related to automated decision-making systems' transparency in order to respond "what information someone using a machine learning system that processes personal should provide? How does this relate to our previous understanding of the questions that one may answer when using a machine learning system and being transparent about it? What is the role of regulators in this regard?"

We thus start with a brief overview of how privacy and data protection evolved in the last one hundred and thirty years or so. Brace ourselves!

### 4.1. From Privacy to Data Protection - A Brief Historical Overview

*International experiences*

When discussing the origins of data protection, the concept of privacy almost inevitably pops up into our minds. One of the first references to a right to privacy was made in the famous 1890 work by Samuel D. Warren and Louis D. Brandeis (1890) "The Right to Privacy". In the article, the authors defend the recognition, in the United States legal system, of the right to privacy as part of the right to one's personality and as a reaction to the increasing intromission in individuals' private sphere by press agents with the use of innovative technologies at the time, especially instantaneous photographs.

Back in those days, the press, according to the authors, overstepped "in every direction the obvious bounds of property and decency" and gossip became a trade and not just the "resource of the idle and the vicious". The authors thus argued that these actions were reaching a new level of intrusion as the increase in newspaper diffusion matched with the use of pictures taken with photo cameras and their publishing in these means of communication. In this sense, the authors defended that the way law could respond to this intromission of private life should be through the recognition and protection of a right to privacy (WARREN; BRANDEIS, 1980, pp. 195-6).

Such a right, they argue, would not arise from contract or special trust, but would consist of a right against the world, or, lending an expression from Thomas Cooley (1888), a "right to be let alone". Such a right would be linked, thus, not to property, but to an individual's sphere of personality (WARREN; BRANDEIS, 1890).

This framing of the right to privacy as a prerogative related to one's personality instead of property brought a new way of seeing the role of personal information in one's existence. Through time, however, as digital technologies developed and the use of computers grew, this subjective way of understanding privacy absorbed additional elements towards a more objective, and sometimes even collective dimension. As such, privacy started to be seen as also encompassing a person's right to keep control over her own information and thus determine the way she aims to develop her own private sphere detached from social control mechanisms. Privacy is thus recognised as a condition to the exercise of other democratic prerogatives such as the freedom of

information and speech, thus going beyond the mere protection of spaces or goods, such as a person's house or communications (RODOTÀ, 1995; DONEDA, 2006).

This enhanced form of comprehending the right to privacy, that goes, therefore, beyond the mere right to be left alone as described both by Colley (1888) and Warren and Brandeis (1890), towards the safeguarding of own's personal information, is the roots for the right to data protection. At this point, it is interesting to highlight how the European Union Agency for Fundamental Rights Agency differentiates both rights in a way that summarises what we have already discussed:

> The right to respect for private life consists of a general prohibition on interference, subject to some public interest criteria that can justify interference in certain cases.
>
> The protection of personal data is viewed as a modern and active right, putting in place a system of checks and balances to protect individuals whenever their personal data are processed (FRA, 2018, p. 19).

In this sense, data protection's main goal would be to provide users control over their data, which results in the power to freely decide for what purpose one's data should be processed (MIRAGEM, 2019). This is the main idea that permeates the right to informational self-determination as described by the German Constitutional Court in a 1983 case that we will discuss below (MAYER-SCHÖNBERGER, 1997). Data protection, thus, extends beyond the mere secrecy of communications, or the safeguarding of private spaces against the intromission of others, reaching the protection of every information that *may* relate to a particular person, up to the moment that individuals should have the control over how this data is processed in our times, permeated by the massive use of digital technologies.

The history behind this worry towards providing control for individuals, as expressed by the German decision, results from a wider debate in the US and Europe from the 1960s onwards, and had a similar root as the one that motivated the drafting of the article by Warren and Brandeis at the end of the nineteenth century: technological developments.

*National databases*

With the enhancement of computers' data storage and processing capacity during the 1960s, several governments, especially in the US and Europe, started to develop policies that aimed at improving the provision of public services through larger personal data gathering and use. By having access to more information about the citizens, policymakers expected to design policies that better reflected the needs of a population, and, hence, many initiatives for creating national, centralised databases proliferated in these territories (DAVIES, 1997).

As Mayer-Schönberger (1997, 222) argues, this came at a good timing, when many nations, especially in Europe, "had just initiated massive social reforms and extended their social-welfare systems". For these reforms to be carried out, more and more information had to be gathered, processed and linked together, and computational technology development was all that was needed by these governments at this time.

In the US, this came in the form of the National Data Center in 1965 (UNITED STATES, 1966). In France, of the *Système Automatisé pour lês Fichiers Administratifs et lê Répertoire de Individus* - SAFARI (BOUCHER, 1974). As a reaction against these initiatives, national debates flourished with worrisome accounts on the potential mass surveillance that could result from this massive data processing and centralisation, leading years later to the approval of legislations such as the US' 1974 Privacy Act and the French 1978 *Loi relative à l'informatique, aux fichiers et aux libertés,* that aimed to provide stronger protection to personal information.

In Germany, differently, a paradigmatic change in the understanding of the right to the protection of personal information came not exactly with a centralised database, but with the Census Act of 1982 (*Volkszählungsgesetz*) (GERMANY, 1983), which allowed authorities to match information from different governmental databases. A broad reaction against the initiative and its wide potential of affecting people's data protection led to the groundbreaking decision of the German Constitutional Court in 1983, which enshrined in the German legal system the right to informational self-determination. According to Doneda (2020), one crucial aspect of the decision was the recognition that there is no such thing as a personal data processing activity which is harmless to privacy. Every information regarding an individual should be protected in

an age where computers advanced and enhanced their capacity of extracting new information from data.

*Privacy and data protection in the books: the first legal provisions and laws*

Privacy and data protection laws proliferated throughout the years, but it is important to note that these were not the first legislations on the theme. The Universal Declaration of Human Rights (UDHR) from 1948 already established a general right to privacy, as well as the European Convention on Human Rights (ECHR) from 1950. With regard to data protection, on its turn, the world's first legislation on the theme had been the Hesse Data Protection Act (*Datenschutzgesetz*), from 1970 (DONEDA, 2020).

It is interesting to note how the social and legal reactions that led to the approval of these legislations were in response to paradigmatic technological advancements and to how these tools were being deployed in larger scale and in different environments. In the case of the 1970s and 1980s initiatives, the social response against the use of computers for what was at the time massive data processing activities came as these devices were no longer deployed strictly in military contexts, but for the processing of information belonging to whole populations by state bureaucracies (MAYER-SCHÖNBERGER, 1997).

Years later, however, the use of computers gradually proliferated, as they were adopted not only by governments, but also by a larger amount of businesses and, ultimately, individuals. It was around this period that data protection came to being recognised as holding a constitutional status, as it happened in the aforementioned 1983 decision of the German Constitutional Court. According to the court's ruling, the right to informational self-determination derived from the general fundamental right to respect for an individual's personality, and should encompass the power of the person to decide for oneself how personal information should be released and used, encompassing as well all phases of information processing, from gathering to transmitting. Further, restrictions to this right could only be made under matters of overriding general interest (GERMANY, 1983).

From these landmarks onwards, much changed in the scope of data protection laws in Europe. As technology continued to develop, especially after the flourishing of the Internet, and private actors started to design their business models based on the massive processing and exploration of personal data collected both online and offline, new protections were included in these rules. From a normative perspective, the requirements for consent to be used as the legal ground for collecting data were increased due to the weakened bargaining position of individuals in face of massive data processors like governments and large technology companies, and some legislations included no-fault based compensation models for individuals' claims. (MAYER-SCHÖNBERGER, 1997)

A pivotal legislation at the European level was the 1995 Data Protection Directive no. 95/46/CE, which embraced most of these evolutions. It established a general prohibition on the processing of sensitive personal data such as those related to the race, religion, political opinions, etc., of an individual, except in a few enumerated cases and purposes. It also included restrictions for the processing of data based on individual's consent, namely the need that consent be informed and expressly given, and determined that data processing activities by both private and public entities should be made under the same level of protection (MAYER-SCHÖNBERGER, 1997). Besides, it contained principles that had already been traced in the Council of Europe's Convention for the Protections of the individuals with regard to the Automatic Processing of Personal Data ("Convention 108") (COUNCIL OF EUROPE, 1981), and the OECD's Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, from 1980 (OECD, 1980). Among these principles, some of them related to the transparency about data processing activities; to the provision of security safeguards; to the limitation of purposes with which specific information should be processed; among others (DONEDA, 2020).

The following decades continued marking important transformations for data protection. In 2000, the European Union proclaimed the Charter of Fundamental Rights (hereinafter "EU Charter") which affirms a specific right to personal data protection in its Article 8, going thus beyond the ECHR's right to privacy. It establishes that data "must be processed fairly for specified purposes and on the basis of the consent of the

person concerned or some other legitimate basis laid down by law" (Article 8.2). The Charter also enshrines rights to access and rectification of data (Article 8.3) and establishes the role of independent authorities in exercising control over the compliance of these obligations.

After the Treaty of Lisbon, in 2009, the Charter, which used to be only a political document signed by the EU Member States, became legally binding with a constitutional status under EU law, and its provisions applicable not only to European institutions at the communitarian level but also to Member States every time they were implementing EU law (FRA, 2018, p. 28).

The year of 2016, however, established a groundbreaking landmark that reflected far beyond the European borders. The approval of the European General Data Protection Regulation (GDPR) reflected important changes in the governance of personal data in Europe, and became the new standard for the subject worldwide.

We will get into more detail regarding the GDPR in the following sections in order to map how its provisions influence the governance of ML systems with a particular focus on transparency. However, it is important to briefly highlight at this point that among the main changes brought by this regulation to the European context was the inclusion of new principles and rights of the data subject. Among them was the addition of a principle of accountability and a right to data portability, in comparison to the Data Protection Directive; an obligation for organisations to implement data protection by design and by default in their internal and external processes; and the obligation to appoint a Data Protection Officer in certain circumstances (FRA, 2018, p. 30).

With such developments, the GDPR was designed aiming to create a barrier against massive personal data processing by both public and private entities, irrespective of their sizes. As the access to data sets and new technologies capable of processing extensive and sensitive amounts of data and of extracting patterns from them (including machine learning) was democratised, now many small companies can also take part in the surveillance capitalism that we talked earlier in the Introduction (ZUBOFF, 2019).

Once we presented this brief introduction to the development of the concepts of privacy and data protection with a focus on the European context and some notes on how privacy was first discussed in the US, it is time now to look at how this theme has evolved in Brazilian law throughout the last decades. As such, a few important legal instruments, including the *Lei Geral de Proteção de Dados* (LGPD, or General Data Protection Law in a translation to English), and the Constitution itself, will be the object of our analysis.

*Data Protection in Brazil*

The first legal provision related to a right to intimacy or privacy can date back, on a constitutional level, to the first Brazilian Constitution, from 1824 (BRASIL, 1824), following the 1822 independence. It enshrines, in its article 179, VII and XXVII, the inviolability of one's home and the secrecy of mail communication, prerogatives that, at least on a normative level persist to the constitutions that followed, even those proclaimed during dictatorial regimes. In 1967, during the military dictatorship, the 1967 Constitution (BRASIL, 1967), the secrecy of communications is extended to the ones held through telephone and telegraph, under article 153, §9th (COLOMBO, 2017).

The current Constitution (BRASIL, 1988), which is into force since 1988 and is a normative landmark after the end of the military dictatorship in Brazil, expanded the reach of privacy protecting provisions at constitutional level. It considers inviolable the private life and intimacy (article 5, X); guarantees the inviolability of communications, including telephonic, telegraphic and, quite dubiously, data (article 5, XII); and institutionalises the constitutional action for *habeas data* (article 5, LXXII), which assures a right of access and correction of personal data about the petitioner contained in records or data banks of government agencies or entities of a public character.

There used to be a considerable debate among Brazilian scholars on whether a right to data protection could be subsumed from the aforementioned provisions (MENDES, 2014; DONEDA, 2020). This discussion came to an end, however, after two important events. The first was the recognition, by the Supreme Court, of the fundamental right status of data protection after declaring the unconstitutionality of a legislation

that obliged telecommunication companies to share consumers' data with the *Instituto Brasileiro de Geografia e Estatística* (Brazilian Institute of Geography and Statistics, IBGE) during the COVID-19 pandemic (MENDES, FONSECA, 2020). The second was after the promulgation of Constitutional Amendment no. 115/2022 (BRASIL, 2022), which included an express provision in article 5 of the Constitution enshrining the status of fundamental right to data protection in Brazil's legal system.

At the legal level, provisions in different acts already assured some limited level of protection to personal data before the entry into force of the LGPD. The *Código de Defesa do Consumidor* (Consumer Defence Code, CDC, Law no. 8,078/1990) (BRASIL, 1990), for instance, established under its article 43 a series of protections for consumers with regard to personal information stored in data sets held in the scope of consumerist relationships. The article and its paragraphs establish, e.g.: a right to have access to information (Article 43, *caput*); an obligation that personal data are correct and that negative information about an individual is stored for no long than five years (Art. 43, §1); an obligation that any inclusion of consumers in registries is informed (Art. 43, §2); a right to rectify data (art. 43, §3), and others.

The provisions of the *Código Civil* (Civil Code, Law no. 10,406/2002) (BRASIL, 2002) also affect the protection of privacy and personal data through provisions on the defence of the rights to one's personality (Article 11) and the inviolability of an individual's private life (Article 21).

Also noteworthy is the *Lei do Cadastro Positivo* (Law for Positive Registry, Law no. 12.414/2011) (BRASIL, 2011a), which regulates the formation of and consultation to databases with information on individuals and legal entities for the formation of so-called positive credit scores, which is data related to people with good credit information. As such, the law provides for rights to the data subject that include:

i. a right to cancel and reopen the registry about oneself (article 5, I);

ii. a right of access to one's personal information(article 5, II);

iii. a right to rectify information(article 5, III);

iv. a right to be informed on the main "elements and criteria" used for the risk analysis (article 5, IV);

v. a right to be informed about the identity of the entity and the manager of the registry, as well as about the storage and purpose of the personal data processing (article 5, V);

vi. a right to request the revision of decisions made exclusively through automated means (article 5, VI);

vii. a right to have personal data used only for the purposes to which they were collected (article 5, VII).

The *Lei de Acesso à Informação* (Access to Information Law, LAI, Law no. 12,527/2011) (BRASIL, 2011b), on its turn, also has provisions on the protection of personal data. LAI regulates the provision of information by public entities in Brazil, and is the result of the country's leadership at the foundation of the Open Government Partnership (OGP), an international coalition with currently 78 Member States dedicated to promote and support the implementation of government policies based on the principles of OpenGov, which involves transparency, public participation, innovation, and accountability (MORAES et al., 2021). LAI provides, under Article 31, that requests for the access to information which includes third-parties' personal data shall be restricted to legally authorised public authorities and its data subject. In this sense, the treatment of personal information must be done in a transparent manner and with respect to intimacy, privacy, honour and image, as well as individual freedoms and guarantees. Exceptions to this restriction are only possible if access to third parties has been provided by law or the data subject has consented to it.

Another law with important effects for the right to data protection was the *Marco Civil da Internet* (Internet Civil Framework, MCI, Law no. 12,965/2014) (BRASIL, 2014). Approved as a response to the developments of the internet, its economic exploration and also to protect individuals against state surveillance, including from foreign actors, the MCI was an internet regulation landmark after the Snowden scandal. As such, it established principles, guarantees, rights and duties of internet users (FRAGOSO, 2019).

These provisions included rights related to privacy and data protection, establishing the (i) inviolability of intimacy and private life (article 7, I); (ii) secrecy of private communications, either shared or stored (art. 7, II, III); (iii) prohibition for internet providers to share parties personal data with third except with free, express and informed consent or in the cases provided for by law (Art. 7o, VII); (iv) right to have clear and complete information about data processing activities and others (BRASIL, 2014; FRAGOSO, 2019).

Besides these fragmented provisions on data protection established through sparse legal instruments, Brazil remained until 2018 without a proper general legislation on the theme. This year, however, marked the approval of the *Lei Geral de Proteção de Dados* (General Data Protection Legislation, LGPD, Law no. 13,709/2018) (BRASIL, 2018), a law with a structure very similar to the European Union's GDPR.

The LGPD marked a new era for digital rights in Brazil, as its provisions allow for a systematic regime of governance and protection for the flow of personal data in our territory. As we will see below in more detail, the LGPD, similarly to the GDPR, applies to any personal data processing activity, through both digital and (structured) physical means establishes, and provides for a set of principles, rights and specific lawful grounds for the processing of data.

Despite LGPD's promises, however, entities from both private and public sectors still struggle to comply to its rules. A study made with 366 companies by a group of consultancies in Brazil identified in 2021 that only 9.8% of the organisations interviewed considered to have between 81% and 100% of its compliance process with the LGPD fulfilled (ALVAREZ & MARSAL *et al*, 2021).

With regard to the public sector, on its turn, the situation can also be worrisome. Security incidents involving public data sets have been common in the last years. Two of the most recent ones involved the Ministry of Health: in 2020, the personal data of at least 16 million suspected or confirmed COVID-19 patients was leaked and exposed on GitHub (ONETRUST, 2020). On 2021, the Ministry's website was hit by attackers who that took several systems down, including one with information about the

COVID-19 national immunisation program and another used to issue digital vaccination certificates (ARAUJO *et al,* 2021).

That without mentioning recent problematic partnerships in which public agencies have been recently involved that included the sharing of personal data of individuals held by the state to private entities for reasons that were criticised by civil society organisations, lawyers and scholars as not clearly related to the public interest (PEREIRA, MORAES, 2022b). These include a norm issued the Brazilian tax revenue agency, *Receita Federal do Brasil* (RFB), which in April 2022 published a norm authorising the *Serviço Federal de Processamento de Dados* (Data Processing Federal Service, SERPRO) to share tax-related data with private companies for "complementing public policies", with no clear information on which public policies would be these and why the data sharing was necessary for such complementation (KNOTH, 2022).

Such incidents show how the Brazilian society still has a long way to go towards a reasonable enforcement of individuals' right to data protection. However, the LGPD is set to be the main normative enabler for this, and its importance is undeniable in this context.

The next few pages will address how the provisions in the LGPD and the GDPR address the regulation of machine learning systems with a focus on transparency. To achieve this goal, we will briefly assess more general issues such as what are their scopes of application, some important definitions such as that of personal data, a general overview of its principles, rights and lawful grounds for data processing specifications, as well as its rules on data processing activities made with the use of automated decision making systems, which are our main interest in this study.

### 4.2. Data Protection Laws in Brazil[3] and Europe

We saw above that the Brazilian data protection legal framework, now represented specially by its general legislation, the LGPD, was profoundly influenced by the European data protection regime. The similarities of both regimes reside, among other aspects, in the scope of application of the acts' provisions, which includes any activities involving the processing of personal data (except for some exceptions which we will see below), in the existence of lawful grounds for processing data, as well as of general principles and rights. The requirements that controllers can only process personal data if there is a legal basis that allow them to do so, that the processing is also in accordance with the principles provided by these laws and that rights are protected, consists of what Mendes et al. (2019, p. 144) call an "ex-ante regulatory rationality".

Other parallels are the creation of centralised authorities responsible for enforcing data protection rules and the existence of diverse obligations for data controllers and processors (MENDES; DONEDA, 2018, pp. 1-2). That despite the fact that only recently the Presidency of Brazil took the initiative of proposing that the authority gain an independency status (ALVES; VALADÃO, 2022), something yet to be assessed by Congress. The independency of the authority is, according to Mendes et al. (2019, p. 146), a basic condition for guaranteeing the effectiveness of the LGPD.

Due to such expressive resemblances, we will throughout this session discuss many topics that are, if not identical, very similar in both legislations. This should not, nevertheless, allow us to be careless about the subtleties that each act holds. An easily detectable difference lies in the more extensive quantity of articles that the GDPR has in comparison to the LGPD (99 and 65, respectively), and to the fact that the latter does not have recitals as its European counterpart does, a feature which would expressively help the interpretation of the provisions.

---

[3] Most of the translations of the LGPD's provisions to English language presented in this work have been adapted from LEMOS, Ronaldo *et al*. Brazilian General Data Protection Law (LGPD, English translation). **IAPP**, [*s. l*], Oct. 2020. Available at: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/. Last accessed: 14 June 2022. The original text of the LGPD can be found at: BRASIL, Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Planalto**. Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Last accessed: 26 June 2022.

Besides that, it is important to note that material dissonances between the two regimes also exist. As we will see, one can find particularities between these two acts in our specific topic of interest, that is to say, data processing activities carried out with the use of automated decision making systems and which information is to be provided in this regard.

This section will thus briefly describe general aspects and some important concepts of these acts in order to provide a broad understanding of the data protection regimes in Brazil and the European Union.

*Material and territorial scopes of application*

The LGPD and the GDPR have similar scopes of application. The LGPD applies to every personal data processing activity, either by public or private entities, irrespective of the means through which this processing takes place (article 1, LGPD), as long as the activity is either (I) carried out in Brazilian territory; (II) has as its goal the offer or supply of goods or services or processes personal data of individuals located in national territory; or (III) when the personal data being processed have been collected in Brazil (article 3, I, II and III, LGPD).

There are, nevertheless, exceptions to the application of the LGPD. According to its article 4, the LGPD does not apply to the processing of personal data that

I. is done by a natural person exclusively for private and non-economic purposes;

II. is done exclusively for journalistic and artistic purposes or, in cases of processing activities with academic purposes, as long as while ensuring, whenever possible, the anonymization of personal data (as provided under articles 7, IV, and 11, II, c, LGPD);

III. when it is done exclusively for purposes of public safety, national defence, state security or activities of investigation and prosecution of criminal offences; or

IV. when the data have their origin outside the national territory and are not the object of communication or shared use of data with Brazilian processing agents or the

object of international transfer of data with another country that is not the country of origin, if the country of origin provides a level of personal data protection adequate to that established in this Law.

It is important to note, however, that even though there is an exception for the cases expressed under the point III above, regarding law enforcement and national defence, the LGPD principles would still apply in data processing activities for these purposes (article 4, §1, LGPD).

One should also observe that the exception for data processing activities within academic purposes is not exactly an exception, although called this way by the LGPD, but a particular purpose which has its own legal ground for lawful processing, which is provided by the aforementioned provisions in LGPD's articles 7, IV and 11, II, c.

The GDPR, on its turn, applies to the processing of personal data, either by public or private entities, wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system (article 2(1), GDPR). Such processing of personal data should be made in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not (article 3(1), GDPR).

Still regarding its territorial scope, the GDPR also applies to the processing of personal data of data subjects who are in the Union even if the controller or processor is not established in the Union. That may happen as long as the data processing is related to (a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or (b) the monitoring of their behaviour as far as their behaviour takes place within the Union (article 3(2), GDPR). Further, it can also apply to the processing of personal data by a controller not established in the Union, but in a place where Member State law applies by virtue of public international law (article 3(3), GDPR).

The exceptions of the GDPR are very similar to the ones provided for by the LGPD. According to its article 2, the GDPR does not apply to the processing of personal data:

a. in the course of an activity which falls outside the scope of Union law;

b. when data is processed by the Member States when carrying out activities which fall within the scope of Chapter 2 of Title V of the TEU, i.e., for foreign policy related activities;

c. by a natural person in the course of a purely personal or household activity; or

d. by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security.

Exceptions under letters "c" and "d" are quite similar to LGPD's exceptions under numbers I and III.

With regard to the exceptions related to data processing activities for journalistic and artistic purposes at the LGPD, there is no direct parallel in the GDPR in the sense that it would not apply at all to these areas. However, its Article 85 establishes that Member States shall by law reconcile the right to data protection "with the right to freedom of expression and information, including processing for journalistic purposes and the purposes of academic, artistic or literary expression" (Article 85(1), GDPR), including through the provision of exemptions or derogations to the GDPR.

*Personal data and data subject*

Having briefly touched upon the scopes of application of the LGPD and the GDPR, it is essential to mention what does *personal data* means. As we have seen, both acts only apply to the processing of such specific types of information, which is why this may be the most important definition of these regimes. Personal data is thus a "threshold concept" for the application of these data protection laws (TOSONI; BYGRAVE, 2020).

The laws have the same definition for personal data, which would be "information regarding an identified or identifiable natural person" (article 5, I, LGPD; Article 4(1),

GDPR). The GDPR's provision, nevertheless, also defines what it means by "identifiable natural person" and gives examples of which information could be considered personal data. In this sense, personal data is that with which

> Article 4(1). (…) one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

The possibility of identification of the individual whose data is concerned — the so-called "data subject" — plays a crucial role here. It is not necessary, for the application of the LGPD and the GDPR, that the person is already directly identified from the data. Even if the person is not yet identified, the mere possibility of this happening through "reasonable means" can also trigger the application of these laws to the data processing activity. That is why both use the term "identifiable" when referring to the individual whose data is being processed, even if the identification is not already given — such as when it is not directly tied to a name — or requires further research to obtain the information necessary to identify the person (FRA, 2018, p. 88). After all, a name is just one of the almost infinite identifiers a person may hold (BORGESIUS, 2016).

In the European law, which influenced the Brazilian in this regard, the reason for such a broad concept of personal data lies in an intention of European policymakers and legislators in extending the GDPR's protection to all information that *may* be linked to an individual, and thus allowing for a greater degree of protection for data subjects (WP29, 2007, p. 4). This logic also includes an intention to address the processing of personal data in the Big Data era, including with the use of machine learning systems, as it happens, for instance, in the profiling of individuals for the purpose of behavioural targeting, such as in the field of advertising, where data is massively used in an attempt to offer ever more personalised products and services to persons (BORGESIUS, 2016).

It is worth noting that the processing of some categories of personal data, the so-called sensitive data (LGPD) or special categories of data (GDPR), receive stronger protection from both acts due to their enhanced potential in allowing for a person to

be discriminated through the knowledge of such information. Under the LGPD, these are personal data concerning racial or ethnic origin, religious belief, political opinion, trade union or religious, philosophical or political organisation membership, data concerning health or sex life, genetic or biometric data, when related to a natural person (Article 5, II).

Under the GDPR, similarly, special categories of data are the ones "revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" (Article 10).

*Data users: controllers and processors*

The main addressees of obligations under both the LGPD and the GDPR are the so-called data controllers and processors[4]. Both of them are entities that carry out personal data processing activities, being the difference between them based on the degree of decision-making capacity regarding the use of such data.

A *data controller* under the LGPD would be a "natural person or legal entity of either public or private law in charge of making the decisions regarding the processing of personal data" (Article 5, VI). Under the GDPR, the controller is defined as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" (Article 4(7)). The main difference between both provisions, thus, lies on the use of the expression "making decisions" by the LGPD as opposed to determining "the purposes and means" by the GDPR.

Differently, the *data processor* would be defined by the LGPD as the "natural person or legal entity of either public or private law that processes personal data on behalf of the controller" (Article 5, VII), while the GDPR's processor would be, very similarly,

4 It is important to note that, in a free translation from Brazilian Portuguese, data processors are called "operators" under the LGPD. Nevertheless, we will address them here as "processors" in order to simplify the analysis of this genre of entities which in both legislations have very similar obligations.

"a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller" (Article 4(8)).

The result of such differentiation is the degree of responsibility that each one of these entities will have for complying with data protection rules, which include allowing for data subjects to exercise their rights (WP29, 2010, p. 4). As data controllers are the ones with the stronger degree of influence regarding the why and how of a data processing activity, it is up to them to demonstrate compliance with the data protection regime (WP29, 2010; EDPS, 2019). It is also worth noting that such obligations apply irrespectively of whether the entity is a public or private entity.

*Principles, rights and legal grounds for lawful processing of personal data*

The LGPD and the GDPR provide principles and rights aimed at guiding the processing of personal data by controllers and processors and at allowing data subjects to gain more information about them and contest the processing if necessary. The principles are referred to by the LGPD under Article 6 and under the GDPR under Article 5, while the rights are expressed under the chapters of number 3 in both acts.

Both laws have also provisions that describe what are the legal grounds over which data can be processed, i.e., under which circumstances can a data controller or processor carry out a personal data processing activity. The LGPD does so under Article 7 (for general personal data) and Article 11 (for sensitive data), while the GDPR provide for them under Articles 6 and 9. These provisions, in the two acts, establish that a personal data processing can be lawful when done so under consent, for the performance of a contract, for complying with regulatory or legal obligations, among others. They are a key element to determine, thus, the lawfulness of a processing.

Important differences between the two laws can be found on the authorising hypothesis for the processing of personal data: while the LGPD contains ten lawful grounds, the GDPR has six in total. The hypothesis that are common to both frameworks are consent, performance of a contract, compliance with a legal obligation, protection of the vital interests of the data subject or of another natural person, performance of a

task carried out in the public interest and legitimate interests. The four additional ones on the LGPD are the data processing by a research body (article 7, IV); (ii) the regular exercise of rights in judicial process (article 7, VI); (iii) health protection (article 7, VIII); and (iv) protection of credit (article 7, X) (MENDES et al, 2019).

Besides the aspects we pointed out, we will not enter into detail with regard to each specific principle, right or legal ground for data processing as this would make us navigate too far from the scope of this study. Nevertheless, we will throughout this work dive deeper into detail with regard to principles such as transparency and rights that include the right to access, as they are closely related to the provision of information regarding activities carried out by ML systems that include the processing of personal data.

*Automated decision-making*

Neither the LGPD nor the GDPR make explicit reference to terms such as artificial intelligence, machine learning or to other techniques usually comprised by the general concept of AI. Instead, they use the term *automated decision-making* to make reference to decisions taken by systems through personal data processing made with some degree of autonomy and independence from humans.

ML systems are mainly automated decision-making technologies. It is the case, for instance, with the content recommendation systems that we explained earlier in this work, that display and rank content on social media platforms without direct intervention from a human being. As such, provisions on this kind of decision, which we may also call conclusion, or output (ALLEN; MASTERS, 2020), may directly affect the regulation of ML systems, reason why they are of particular importance for this work.

Neither the LGPD or the GDPR provide a straight forward definition on what automated decision-making means. Nevertheless, both have specific rights and obligations that apply to them, even if in different degrees of protection, as one can conclude from an analysis of LGPD's Article 20 and GDPR's Articles 13, 14, 15 and 22.

To clarify the concept and the consequences of automated decision-making, the WP29 issued the Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. In the work, the group affirmed that, under the GDPR, "[s]olely automated decision-making is the ability to make decisions by technological means without human involvement" (WP29, 2018, p. 8). A similar wording regarding "solely automated decision-making" is used under LGPD's Article 20[5]. This type of decisions, according to the Working Party, can be based on any type of data, including data provided directly by the individuals concerned; data observed about the individuals; and those derived or inferred data such as a profile of the individual that has already been created (WP29, 2018, p. 8).

*Human involvement in automated decision-making*

Under both the GDPR and the LGPD, understanding the degree of involvement of a human being in the decision-making process is important to assess whether specific rights and obligations, mostly related to the provision of information and human revision, are provided.

According to the WP29, to qualify as human involvement, any oversight of the decision should be "meaningful, rather than just a token gesture" and "should be carried out by someone who has the authority and competence to change the decision" if necessary (WP29, 2018, p. 21). This leads us to understand that to qualify as an involvement, it is necessary that there is an approval of the system's output by a person, to a degree that involves a qualified assessment of whether the decision is indeed correct or makes sense. In this sense, the acceptance without meaningfully questioning the conclusion of the system should be enough to qualify it as a solely automated decision-making.

---

5 "Art. 20. *O titular dos dados tem direito a solicitar a revisão de* ***decisões tomadas unicamente com base em tratamento automatizado de dados pessoais*** *que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade."* In English: "The data subject has the right to request for the review of **decisions made solely based on automated processing of personal data** affecting her/his interests, including decisions intended to define her/his personal, professional, consumer and credit profile, or aspects of her/his personality" (my emphasis). Translation by LEMOS, Ronaldo *et al*. Brazilian General Data Protection Law (LGPD, English translation). **IAPP**, [*s. l*], Oct. 2020. Available at: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/. Last accessed: 14 June 2022.

Under GDPR's Article 22(1), data subjects "shall have the right not to be subject to a decision based *solely* on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (our emphasis).

Such provision shall not apply, however, if the decision: "(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent" (Article 22(2)(a)(b)(c)). Nevertheless, the GDPR provides that, in these cases, further protection must be provided in the data processing, which includes a right to human intervention, according to Article 22(3).

The protection provided by LGPD is weaker than in its counterpart, as it does not provide for a general prohibition on decisions based solely on automated processing of personal data. Further, with regard to reviews of the decision, it provides merely a right to request the review of "decisions made *solely* based on automated processing of personal data affecting her/his interests" (Article 20). Such a right does not specify that the review should be made by a natural person.

One should note, however, that this was not always the case. The original wording of the LGPD obliged data controllers to provide for human revision, but the text was afterwards vetoed by President Jair Bolsonaro, who argued that this would inhibit the development of startups and the assessment of credit-worthiness (BRASIL, 2019).

Having assessed key provisions in both LGPD and GDPR, we now turn to an analysis of how their provisions address the access to information regarding the use of automated decision-making systems or, more specifically, machine learning.

### 4.3. Machine Learning Transparency under the LGPD and the GDPR

In Chapter 3, we discussed whether and to what extent providing for greater transparency in machine learning systems could be an effective tool to better understand

their impacts to society, the environment or any other field humans may exert influence through them. We saw that, despite not a panacea, transparency can indeed be a tool to gain insight into how the use of these systems is changing the world we inhabit, for better or for worse, and that we should learn how to pose the right questions in order to access information about them that is effective to achieve the goals we may have. Having said that, it is now the moment to understand whether and how this theme is addressed by data protection laws in Brazil and the European Union.

In the last section, we saw that data protection regimes in these jurisdictions have specific obligations and rights that apply when automated decisions are made based on the processing of personal data. As machine learning systems are mostly applied for carrying out such automated decision-making (ADM) processes, these rules are directly applicable to those making use of these technologies. For this reason, they are of particular interest for us in this study, especially considering that some of these provisions also extend to the delivery of information regarding the operation of these systems and, consequently, their transparency.

A large discussion has taken place in the last years over whether there is or not a right to explainability in these laws, mostly focusing on specific provisions related to ADM regulation. In the European Union, where these discussions came earlier in Brazil, Goodman and Flaxman (2016) published one of the first articles that affirmed that there was a right to explanation in the GDPR by setting provisions for the necessity in providing information for a data subject regarding a personal data processing made through automated decisions.

Wachter, Mittelstadt and Floridi (2017) reacted affirming that one could not extract such a right from the regulation, but merely a right to information, as there would be no right for explanations of single algorithmic-based outputs, but merely a “right to information” of the general logic of a system.

From then on, many other scholars such as Selbst and Powles (2017), Malgieri and Commandé (2017) and Kaminsky (2019) have been dwelling on this discussion. Others, such as Edwards and Veale (2017), as we showed earlier, highlight that there is a

"transparency fallacy", while Annany and Crawford (2017) drew attention to how transparency can be used as a tool for further opacity.

In Brazil, inspired by the European debate, Leite (2018), Souza, Perroni and Magrani (2020), as well as Lima and Sá (2020), have also delve into this debate, reaching a similar conclusion that in some situations the LGPD may provide for a right to explanation.

This work will shed light on their arguments above. However, it is worth anticipating that its aim is no more to discuss whether or not such a so-called right to explanation exist in these jurisdictions, but instead *how* should information regarding personal data processing carried out by ML systems, be it regarding the system as a whole, a specific output or what we have been calling the system's environment or context.

Kaminsky (2019, p. 209) makes an important statement when she says that "in the right to explanation debate, the centrality of transparency to the GDPR has gotten lost". Following not only her ideas, but also those of others, I consider that the LGPD and GDPR both provide for rights and obligations regarding the provision of information that go beyond the discussion related to the right to explanation, another reason that adds to the other ones that made us focus on the term *transparency* instead of explainability in this work. Now it is our duty to understand how to operationalise these rights and obligations.

To do so, before we look to transparency provisions specifically referred to ADM, we should rather give a step back, to the principles and rights provided by the LGPD and the GDPR that apply for every personal data processing, and not just those carried out through automated decision-making. As we will see, not only there are specific transparency principles in both laws, but such transparency is necessary for ensuring the mere capacity of the data controller to prove compliance with the LGPD's and GDPR's principles. This work will call this "general transparency", as, in our specific context, it relates to transparency obligations that go beyond the use of machine learning systems.

### *4.3.1. General transparency in personal data processing*

We saw above, even without entering into much detail, that the LGPD and the GDPR have principles that apply to every personal data processing under their scope. Such principles are conditions for the lawfulness of these operations and, although sometimes framed differently in the two regimes, one can say that they hold significant similarities.

The table below presents a comparison between the wordings of the principles in the two laws[6]:

| LGPD | GDPR |
|---|---|
| Article 6. Activities of processing of personal data shall be done in good faith and be subject to the following principles: | Article 5.1. Personal data shall be: |
| I – **purpose**: processing done for legitimate, specific and explicit purposes of which the data subject is informed, with no possibility of subsequent processing that is incompatible with these purposes; | (b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('**purpose limitation**') |

6 English translations of the LGPD's provisions have all been adapted from LEMOS, Ronaldo *et al*. Brazilian General Data Protection Law (LGPD, English translation). **IAPP**, [*s. l*], Oct. 2020. Available at: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/. Last accessed: 14 June 2022. For the original text of the LGPD, please refer to: BRASIL, Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Planalto**. Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Last accessed: 26 June 2022.

| LGPD | GDPR |
| --- | --- |
| II – **adequacy**: compatibility of the processing with the purposes communicated to the data subject, in accordance with the context of the processing;<br><br>III - **necessity**: limitation of the processing to the minimum necessary to achieve its purposes, covering data that are relevant, proportional and non-excessive in relation to the purposes of the data processing; | (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('**data minimisation**'); |
| IV – **free access**: guarantee to the data subjects of facilitated and free of charge consultation about the form and duration of the processing, as well as about the integrity of their personal data; | *No direct counterpart among the GDPR's principles.* |
| V – **quality of the data**: guarantee to the data subjects of the accuracy, clarity, relevancy and updating of the data, in accordance with the need and for achieving the purpose of the processing; | (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('**accuracy**'); |
| VI – **transparency**: guarantee to the data subjects of clear, precise and easily accessible information about the carrying out of the processing and the respective processing agents, subject to commercial and industrial secrecy; | (a) processed lawfully, fairly and in a *transparent* manner in relation to the data subject ('lawfulness, fairness and **transparency**'); |

| LGPD | GDPR |
|---|---|
| VII – **security**: use of technical and administrative measures which are able to protect personal data from unauthorized accesses and accidental or unlawful situations of destruction, loss, alteration, communication or dissemination; | (f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (**'integrity and confidentiality**'). |
| VIII – **prevention**: adoption of measures to prevent the occurrence of damages due to the processing of personal data; | *No direct counterpart among the GDPR's principles.* |
| IX – **nondiscrimination**: impossibility of carrying out the processing for unlawful or abusive discriminatory purposes; | (a) processed lawfully, **fairly** and in a transparent manner in relation to the data subject ('lawfulness, **fairness** and transparency'); |
| X – **accountability**: demonstration, by the data processing agent, of the adoption of measures which are efficient and capable of proving the compliance with the rules of personal data protection, including the efficacy of such measures. | 2. The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('**accountability**'). |

Table 2 - Principles in the LGPD and the GDPR

With the table above, we intend to trace parallels (even if winding ones) between the principles provided by the LGPD and the GDPR in order to facilitate assessment of their content and identify similarities and differences between them. For instance, one can already identify that there are no counterparts in the GDPR for the LGPD's principles on free access and prevention. Nevertheless, although not having the same normative framing of a principle, the right to access provided in GDPR's Article 13 (and which has a similar counterpart under LGPD's Article 18, II) may deliver some-

thing close to what the LGPD's principle on free access does. Similarly, the lack of a prevention principle can to some degree be mitigated by provisions on data protection by design and default in the GDPR (Article 20), which are also present, with particularities, under LGPD's Chapter VII.

Other disparities can be found in the wording of principles that are similar among both laws, but that hold specific differences. An example lies in the LGPD's data quality principle and GDPR's accuracy. Although the two of them provide for guaranteeing that data is accurate and up to date, the LGPD includes the notions of clarity and relevance of the personal information to the purpose of the processing. The GDPR does not have a similar parallel in its specific accuracy principle, but, on the other hand, it adds that inaccurate data should be erased and rectified without delay.

Further, it is possible to find some similarity in the notion of "fairness" of the GDPR and of "nondiscrimination" in the LGPD. According to the European Data Protection Board, "[f]airness is an overarching principle which requires that personal data should not be processed in a way that is unjustifiably detrimental, unlawfully discriminatory, unexpected or misleading to the data subject" (EDPB, 2019, p. 17-8). One might also find a path to trace such a parallel, despite the lack of a direct, precise translation of "fairness" to Portuguese, when taking into consideration the meaning proposed by the Cambridge Dictionary, which describes fairness as "the quality of treating people equally or in a way that is right or reasonable" (CAMBRIDGE, 2022).

It is worth noting that the parallels herein traced between principles in the LGPD and the GDPR are not unanimous among scholars. Mendes et al. (2019, p. 159), for instance, consider that the GDPR principles of integrity and confidentiality would find a counterpart not in the LGPD's principle of security, but instead of data quality. They also do not see a direct parallel between the LGPD's nondiscrimination principle and the GDPR's fairness principle.

Despite such divergencies, it is not our focus here to address what is the content of each specific principle in the LGPD and the GDPR, but to assess whether these provi-

sions may have an impact in the transparency of ML systems.[7] Our analysis will thus be made accordingly.

We start with the transparency principles in these laws. Under LGPD's Article 6, VI, the transparency principle establishes that the data subject should have "clear, precise and easily accessible information about the carrying out of the processing and the respective processing agents". The GDPR, in more general terms, expresses in its Article 5(1)(a) that personal data shall be "processed (…) in a transparent manner".

Döhmann (2020), when commenting the GDPR's principle, affirms that it contains two main elements. First, it aims to allow for a data subject to gain control over how his or her personal data is being used. Second, that those responsible for the data processing should also constantly verify the lawfulness of the data treatment, including through a duty to document that can be identified through a reading of other rights provided by the GDPR, which may be found in Articles 12, 13, 14, 15 and 22.

In this sense, transparency is also an indispensable requirement for a data controller to prove compliance with the other principles of the GDPR. That includes proving that a specific processing of personal data, resulting or not from automated decisions, have not been unfair or based in inaccurate data, for instance. In order to understand, for instance, whether a data controller making use of an ML system is fulfilling the data quality or accuracy principles, a data subject could make use of the transparency principle, alongside with rights of access to data that are expressed in other articles of the GDPR and the LGPD, to receive information on that, as we will see.

And speaking of the LGPD, its transparency principle follows the same direction when providing that "clear, precise and easily accessible information about the carrying out of the processing and the respective processing agents" should be guaranteed. Such a wording can be even more precise than the GDPR's, as it specifies that the information to be provided shall be, in other words, meaningful for the data subject.

This might be interpreted as a way to make communications regarding data processing activities comprehensible, as Kaminsky (2019, p. 212) does when discussing the

[7] For an in-depth comparison between the LGPD and the GDPR based on the neo-institutional theory of law, please refer to Mendes et al. (2019).

GDPR's Article 12, which establishes that information regarding the processing of data through ADM be "concise, transparent, intelligible and easily accessible form, using clear and plain language". As the author also mentions, this can also be seen as a way to avoid the strategic opacity highlighted by Annany and Crawford (2016), as we discussed in Section 3.3.

Based on the above, we may conclude that the principle of transparency is, in both laws, an umbrella obligation that puts transparency as a general rule for data controllers to apply in every data processing, irrespectively of being made through automated decision-making means, and that the information is sufficiently meaningful. It leaves, on the other hand, open questions. What specific information should be considered as suitable to fulfil the requirements of these laws? What should be considered, looking e.g. into the LGPD's transparency principle and the aspects highlighted in the last paragraph of Article 12 of the GDPR, "clear, precise and easily accessible", or "concise, transparent, intelligible and easily accessible" for the Brazilian and European acts, respectively?

We can already anticipate that not much detail is provided by any of the regimes, reason why so many scholars have still been discussing this theme. That is why we now turn to provisions that specifically address the transparency of automated decisions and how they have been interpreted in Brazil and the EU.

### *4.3.2. Automated decision-making transparency*

Considering, as we mentioned above, that the Brazilian debate on a right to explainability has been mostly inspired by the one made in the scope of the GDPR, we will first focus on how the European debate developed, and then turn towards the one held under the LGPD. With this approach, we hope to set a better understanding of how the discussion has evolved during the last years in order to define our own thoughts on the matter.

*Automated decision-making transparency in the GDPR*

The GDPR sets specific provisions regarding automated decision-making transparency. It does so by creating obligations for data controllers to provide information about the processing of personal data through automated decision-making systems, as established in Articles 13(2)(f), 14(2)(g) and 15(1)(h) (ASGHARI *et al*, 2021).

Reading these GDPR provisions, one can identify that the following information should be provided by the data controller to the data subject in case there is a processing of personal data by an automated decision-making system:

1. The existence of automated decision-making;
2. "Meaningful information about the logic involved" in the automated decision;
3. Significance and the envisaged consequences of such processing for the data subject.

It is worth noting that such obligations apply in situations where personal data are collected directly from the data subject (Article 13(2)(f)) or not (Article 14(2)(g)). Further, Article 15's rules are presented within a general right of access of the data subject, which encompasses a right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, as well as access to the personal data and further information (Article 15(1)(h)).

Under the GDPR, all such information shall be provided proactively by the data controller, without the need of request from the data subject. It is worth noting that the GDPR aims to protect data subjects especially in regard to automated decisions that are carried out for profiling purposes based on the processing of sensitive data, something which would express a major risk for one's data protection.

The GDPR is very strict as to which automated decisions shall be allowed when dealing with personal data processing. Under its recital 71, it describes that evaluating, through automated decision-making means, the personal aspects relating to a natural person where it produces legal effects, shall only be made where expressly authorised by Union or Member State law. Such processing shall always be subject to suitable safeguards, "which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an

explanation of the decision reached after such assessment and to challenge the decision" (Recital 71, GDPR).

Although the GDPR highlights the sensitiveness of automated decisions about an individual and how they may impact the exercise of the rights of this person, the Regulation is not very specific as to which exact information shall be provided regarding the processing (and the system used) and how. There is also disagreement as to whether the regulation requires controllers to explain merely the underlying logic of these systems or whether that would also apply to specific decisions automatically made, and, in these cases, which of such outputs should be explained.

For instance, Wachter, Mittelstadt and Floridi (2017) argue that the GDPR establishes not a right to explanation of individual decisions, but instead a mere "right to be informed". Such a right would encompass information regarding solely the general logic involved in the system as well as the significance and the envisaged consequences of automated decision-making systems. If we apply the taxonomy provided in Chapter 3 regarding different methods for system explainability, such an interpretation would lead for the understanding that the GDPR obliges controllers to provide only global model interpretability, and not information regarding how specific decisions were taken (local model interpretability). The authors' reasoning results mainly from the idea that the term *logic involved* in the automated decision refers only to general information about the system, and not to how the specific decision has been made.

On the opposite direction, Selbst and Powles (2017) interpret that the GDPR should be read as establishing a right to explainability to the extent that the data subject should be provided sufficient information in order to exercise his/her rights. In other words, every time the rights of a data subject is put at risk because of the processing of personal data through automated decision-making, any explanation which is necessary to assess whether there is indeed a violation of the GDPR shall be provided. That would include, for instance, means for identifying whether a specific automated decision has been biased or not. Therefore, there would not be any *a priori* restriction about the explainability method to be applied, as long as it provided sufficient information for the exercise of one's data protection rights (SELBST & POWLES, 2017).

Selbst and Powles' (2017) interpretation for the GDPR concerns a more systematic approach towards the regulation. When addressing data processing by automated decision-making systems, Recital 38 of the GDPR states that "[p]rofiling that results in discrimination against natural persons on the basis of personal data which are by their nature particularly sensitive in relation to fundamental rights and freedoms should be prohibited under the conditions laid down in Articles 21 and 52 of the Charter." Article 11, GDPR, complements such reasoning by positing that "[p]rofiling that results in discrimination against natural persons on the basis of special categories of personal data referred to in Article 10 shall be prohibited, in accordance with Union law".

To illustrate their view, we may consider that it would be quite challenging to assess whether a specific automated decision based on profiling activities might have been biased without analysing the specific output of the system, and not just the underlying logic of the machine learning model as a whole. In this sense, it would seem more appropriate to argue that the GDPR addresses a right to explainability at least in profiling cases, and especially those carried out through the processing of sensitive data.

However, as we mentioned above, and agreeing with Selbst and Powels (2017) and with Kaminski (2019), discussing whether there is a right to information or explainability can be a distraction in the process of understanding the transparency provisions in these laws. That applies also to the LGPD, as we will see further.

The general transparency systematic of the GDPR is about allowing for the controller to prove compliance with the Regulation and for a data subject to gain control over how his or her personal data is being processed in order to exercise the rights provided by the act. That is why Kaminski (2019, p. 213) affirms that communication to individuals about algorithmic decision-making must be "simultaneously understandable (…), meaningful, and actionable", so that someone who is subject to algorithmic decision-making through the processing of personal data can invoke one's rights (KAMINSKI, 2019, p. 215).

This would be the path for the "qualified transparency" that Pasquale (2015) advocates for, in a way that transparency should allow for the targeting of revelations related to automated decisions that have "different degrees of depth and scope aimed at

different recipients". Transparency would be thus not limited to revelations to individuals, but also to regulators, third parties, the public and within the own company, so there is enough internal oversight so that data is not misused. And each of these actors may have access to different types of information in order to make the GDPR's enforcement effective (KAMINSKI, 2019, p. 210-1).

In this sense, the assessment of what does the GDPR mean by "logic" and "meaningful information" should be made in a more systemic manner, taking into consideration the principles and the general protection frame of the Regulation. In accordance to the discussion we had above, in Chapter 3, the GDPR's structure prescribes that the information to be provided regarding automated decision-making should encompass all the information that is necessary for a person to exercise her rights. This does not need to be *every* information about the decision, but the disclosure needs to be effective to the extent that it allows for someone to take action based on it. When reaching this point, the specific requirements on information provision regarding automated decisions meet the general transparency principle of the GDPR (KAMINSKI, 2019).

That is why Asghari *et* al (2021), after noting that neither "logic" nor "meaningful information" are legally defined within the GDPR, affirm that although a complete disclosure of an algorithm's source code, for instance, is not always necessary to address its impacts, such disclosure may be required if this is the only way for the data subject to notice if his or her rights have been violated and take action based on it. As the authors put it, the information provided need to allow an "actual gain in knowledge" (ASGHARI, 2021, p. 9).

Nevertheless, a note should be made with regard to the potential limitations of GDPR's Article 22(1). We saw that this provision, which establishes rules for automated decision-making to comply with the Regulation, limits its scope to decisions based *solely* on automated decision-making "which produces legal effects concerning him or her or similarly significantly affects him or her".

When we read Articles 13(2)(f), 14(2)(g) and 15(1)(h), we can see that all of them have an identical wording, which is that, for providing for fair and transparent data processing, one should provide for the following information:

> 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. (Articles 13(2)(f)/14(2)(g)/15(1)(h), our emphasis).

From our emphasis, we can see that all of them make reference to Article 22(1), which has the restriction we mentioned, towards applying only to decisions producing legal or similarly significant effects on individuals. Nevertheless, in an opposite direction to what authors like Wachter et al. (2017) argue, this should not be seen as a way to narrow the scope of a right of individuals to have transparency or explainability when affected by automated decision-making, even when the interests affected are not as "significant" as legal ones or when decisions are not made under "solely" based automated decision-making.

Instead, the idea that transparency in the GDPR should be seen as enforceable towards any system as a requisite to demonstrate compliance or to allow for the access to rights. This matches also with the basis of the regulation's rights-based approach that we referenced above (FRA, 2018). We can argue that not only based on the same arguments regarding the transparency principle that we mentioned earlier, but also on the fact that Articles 13(2)(f), 14(2)(g) and 15(1)(h) affirm that meaningful information should be provided "at least in those cases" mentioned by Article 22(1). This notion would thus allow to the application of transparency provisions far beyond decisions producing legal effects.

Having looked at the GDPR's provisions, and how transparency should be seen as a general obligation that applies to the whole law, and that debating whether or not there is a right to explanation in it  might distract us from understanding the broader —and deeper — scope of the transparency principle, it is now time to look at the LGPD and see whether this logic is also found under its rules.

*LGPD's transparency framework*

The LGPD also addresses a right to request information regarding the functioning of automated decision-making systems, including those intended to define one's person-

al, professional, consumer or credit profile or aspects of one's personality under Article 20. Such a right can also be seen as a form of materialising and complementing the transparency principle provided by Article 6, VI, which establishes, as we saw above, that clear, precise and easily accessible information shall be provided regarding every personal data processing under the LGPD's scope (MONTEIRO, 2018). Along with LGPD's Article 9, these provisions provide for a broader *transparency framework* for the legislation.

However, differently from the GDPR, there is no specific obligation on the LGPD for the controller to provide such information proactively, but only under request from the data subject. It is worth noting, in addition, that although there is a rule that allows for data subjects to ask for revisions of decisions made by automated systems, there is no specific requirement that this revision should be made by a human agent.

According to Article 20, §1º, LGPD, "[w]henever requested to do so, the controller shall provide clear and adequate information regarding the criteria and procedures used for an automated decision, subject to commercial and industrial secrecy". In case the controller does not provide information, even if by arguing that the disclosure would infringe commercial and industrial secrecy, the LGPD establishes that the national authority may carry out an audit to verify specific discriminatory aspects in the automated processing of personal data.

In this sense, the LGPD leaves enough room for systems to be audited and for controllers to be obliged to provide explanations under the request of a data subject. As is the case in the GDPR, although there is no explicit obligation that controllers will have to disclose information regarding specific decisions made by their automated systems, the same logic that the disclosure by the data controller should be sufficient for the latter to prove, or for a data subject to understand, whether the data processing has been made lawfully, is present.

In any case, one may identify that the LGPD's wording, once again similar to the GDPR, presents a considerable degree of uncertainty as to how and for what purposes information should be provided, particularly regarding what it means by the "criteria and procedures used for an automated decision". The lack of a legal definition for this

expression is similar as to what happens in the GDPR with "meaningful information about the logic involved" in Articles 13(2)(f), 14(2)(g) and 15(1)(h).

Here, the disclosure of information related to automated decisions in the LGPD should be interpreted similarly as we did with the GDPR, in the sense that the information provided should be enough for a subject, including an auditor, regulator, researcher or the data subject, to sufficiently understand the personal data processing and take action based on that (SOUZA; PERRONE; MAGRANI, 2021).

To support this argument, which once again takes into consideration a global interpretation of the law, and not just a strict reading of the provisions that relate to automated decision-making in Article 20 of the LGPD, we may look at what other provisions related to information access describe as necessary information to be provided by the data controller.

LGPD's Article 9, *caput*, I and II, establishes that:

> Art. 9. The data subject has the right to facilitated access to information concerning the processing of her/his data, which much be made available in a clear, adequate and ostensible manner, concerning, among other characteristics provided in regulation for complying with the principle of free access:
> I – the specific purpose of the processing;
> II – the type and duration of the processing, being observed commercial and industrial secrecy; (…)[8]

From this provision, we may conclude that, in general, data subjects have a right to access information related to the specific purpose and of the type and duration of any data processing that encompasses his or her personal data. We may add to these information the requirements from Article 18, I and II, that the data subject also has the right to obtain from the controller the confirmation of the existence of the data processing and access to the personal data being used by the controller.

---

[8] As translated by LEMOS, Ronaldo *et al*. Brazilian General Data Protection Law (LGPD, English translation). **IAPP**, [*s. l*], Oct. 2020. Available at: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/. Last accessed: 14 June 2022. Original wording, in Portuguese: *Art. 9º O titular tem direito ao acesso facilitado às informações sobre o tratamento de seus dados, que deverão ser disponibilizadas de forma clara, adequada e ostensiva acerca de, entre outras características previstas em regulamentação para o atendimento do princípio do livre acesso: I - finalidade específica do tratamento; II - forma e duração do tratamento, observados os segredos comercial e industrial.*

With that in mind, it is suffice to say that, when regarding automated decisions, a data controller should disclose at least the information highlighted by the above mentioned provisions, which include:

a. the specific purpose of the processing;
b. the type and duration of the processing, being observed commercial and industrial secrecy;
c. confirmation of the existence of the data processing;
d. what personal data is being used by the controller.

Nevertheless, the disclosure shall not be restricted to these information unless they suffice to allow for, as we mentioned earlier, the data subject to exercise his or her rights and for the controller to prove compliance to the LGPD. When that is not the case, every necessary information to reach this objective should be provided in order to comply with the transparency principle.

This is why, similarly to what we discussed concerning the GDPR, it also makes no sense to discuss whether there is or not a right to explainability regarding specific decisions under the LGPD. Every information that is necessary, meaningful and sufficient, in a specific case, to reveal whether the controller is complying with the law or not shall be provided.

Finally, when there are commercial secrets involved, auditing by the National Data Protection Authority (ANPD) should be carried out, according to the wording of Article 20, §1, mentioned above (SOUZA; PERRONE; MAGRANI, 2021). This provision can be read jointly with Article 55-J, XVI, of the LGPD, which includes among the ANPD's prerogatives the power to "carry out audits, or to determine their occurrence regarding the processing of personal data carried out by processing agents, including public authorities".

As such, under both the LGPD and the GDPR, the amount of information that will have to be provided by the data controller concerning the data processing will thus have to be assessed on a case by case basis, and the analysis of whether the informa-

tion is sufficient depends on to what degree it allows for the data subject to exercise rights.

Before moving forward, however, we should repeat here the same exercise we did with the GDPR to understand the potential limitations of LGPD's Article 20. If we look at the *caput* of Article 20 and, once again, to its §1º, these provisions will read as follows:

> Art. 20. The data subject has the right to request for the review of decisions made **solely** based on automated processing of personal data **affecting her/his interests**, including decisions intended to define her/his personal, professional, consumer and credit profile, or aspects of her/his personality. (new wording given by Law No. 13,853/2019) (our emphasis)
> §1 Whenever requested to do so, the controller shall provide clear and adequate information regarding the criteria and procedures used for an automated decision, subject to commercial and industrial secrecy.[9]

Reading through Article 20, *caput*, one can identify that the review of decisions can be made of decisions "solely" based on automated processing of personal data that affects a data subject's "interests".

From the outset, this could be seen as a far less narrow provision in comparison with GDPR's Article 22(1). It does not restrict the review of solely automated decisions only to those having legal or similarly significant effects on data subjects, but instead to decisions that affect "interests". Nevertheless, our question is here whether explanations/transparency mechanisms should also be seen as limited to decisions solely based on automated decisions that affect one's interests.

Through a strict interpretation of Article 20, §1, one may say that not necessarily. After all, the provisions establishes a rule to provide "clear and adequate information regarding the criteria and procedures used for **an** automated decision" (our emphasis),

---

[9] As translated by LEMOS, Ronaldo *et al*. Brazilian General Data Protection Law (LGPD, English translation). **IAPP**, [*s. l*], Oct. 2020. Available at: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/. Last accessed: 14 June 2022. Original wording, in Portuguese: *Art. 20. O titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade. (Redação dada pela Lei nº 13.853, de 2019). §1º O controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial.*

without repeating the wording related to “solely” or “interests”. But these are not solid grounds for asserting this idea. Instead, once again, we believe that this should not be seen as a limitation when the compliance with the LGPD is in question, considering the general focus on transparency as a principle that exists in the law. With this notion, it is important to note that we not exactly go against what, for instance, Monteiro (2018, p. 20) affirms when interpreting this provision, that the restriction on “interests” should be observed. After all, a potential lack of compliance with the LGPD or barriers to the exercise of rights can hardly be something that would not be considered an interest of the data subject being affected.

### 4.4. Beyond the books: open questions

These conclusions leave us with more open questions. The first ones relate to what the LGPD and the GDPR do not provide for. We mentioned above that many important impacts caused by the use of ML systems relate to issues arising from the processing of personal data. However, ML pitfalls are not related only to data protection.

As we mentioned, the deployment of ML applications to control individuals in working spaces has been creating physical and mental issues (IRANI, 2016). The development of ML systems and the manufacturing of devices may be having an intense environmental and humanitarian cost that we cannot even measure yet (CRAWFORD, 2021). Lack of digital literacy might be allowing for people to trust these systems blindly and not understand how they are impacting their access to information (BUCHER, 2018). Group privacy issues posed by ML systems processing anonymised data may also have a hard time in being protected through data protection laws (TAYLOR, 2017).

These are only a few examples for us to have in mind that regulating ML goes far beyond data protection, and that, for this reason, not every problem related to opacity raised in the previous chapters will be solved through an application of the LGPD or the GDPR.

Moreover, once transparency in automated decisions — and now we come back to our original research theme, machine learning systems — is contextual and thus highly dependant on case by case analysis (ASGHARI et al., 2021), we need to start asking ourselves how should the enforcement of these provisions take place. Answering this question is a complex task not only due to the broad and quite indefinite scope of the transparency provisions we assessed, but also due to the complex structure of enforcement incentives that the LGPD provides and the various actors that can be involved in this process. That is why we now turn to an analysis of what regulatory strategies may be applied to enforce these transparency provisions.

It is important to note that, from now on, our focus will straighten towards the Brazilian legal system and thus the LGPD. What happens in the European Union will serve in certain occasions as a source of ideas and inspiration, but the next chapter no longer aims to conduct a comparative analysis between what happens (or should happen) in Brazil and Europe with regard to regulatory strategies.

In this sense, when we think of enforcement and oversight, it is reasonable that the first thoughts that arise are driven towards data protection authorities, which in Brazil is, as we mentioned previously, the *Autoridade Nacional de Proteção de Dados,* ANPD. It is therefore its role in enforcing transparency rules that most interests us at this moment. Nevertheless, as we will see, other stakeholders may have a significant importance in this regard. Courts, other regulators, civil society organisations, auditors, national councils, standardisation bodies, data controllers, operators, data subjects, alone or in groups, among many others, have also a role to play in the oversight of ML applications.

In the next chapter, we will thus discuss how enforcement of ML transparency can occur in Brazil under the guidance of the responsive regulation theory.

## 5. REGULATING MACHINE LEARNING SYSTEMS, ENFORCING TRANSPARENCY: AN ANALYSIS THROUGH THE LENS OF THE RESPONSIVE REGULATION

From the previous pages, we can identify many features of machine learning systems that make this technology particularly challenging to regulate, despite its known impacts on society.

A myriad of factors contribute to this. They include, e.g., the complexity of its functioning; the sometimes non obvious nature of its outputs and the difficulty in explaining the details of how outputs are made; ML's transversal nature, which allows it to be applied in fields that range from healthcare to law enforcement, that are subject to distinct rules and oversight authorities; its potential to develop, in the future, towards directions that are hard to predict; the narratives of technosolutionism and inevitability that run across society regarding machine learning and that are strongly sponsored by the developers and deployers of these technologies; the complex chain of actors, materials and information involved in the development, deployment and disposal of ML systems; its main raw material, data, does not seem to raise any scarcity issue (BLACK; MURRAY, 2019), but instead seem to ever increase in quantity.

These are some of the main elements that make machine learning, and AI in general, a challenging range of technologies to regulate, and have motivated the creation of different narratives on which kind of ethical or legal rules should be designed to deal with ML's risks.

One of these narratives rely on the notion that it is yet too early to regulate artificial intelligence systems — few, if any, of the regulatory initiatives make reference to the term machine learning as its main object — especially if such regulation would hinder innovation (GURKAYNAK; YILMAZ; HAKSEVER, 2016).

Another approach with a similar perspective against regulation consists of the argument that the market itself would be responsible for regulating AI based strictly on soft law mechanisms and codes of practice (BLACK, MURRAY, 2019). This trend is closely linked to the last years' trend of designing ethical principles to guide AI de-

velopment and deployment, which, as we showed earlier in Chapter 3, have proliferated among dozens of proposals authored by companies, civil society organisations, think tanks and governments (FJELD et al., 2020), including even the Vatican (VATICAN, 2020).

This wave of "ethics washing" and the argument that the translation of these principles into soft law and codes of practice should be sufficient for governing AI systems has been seen with skepticism by scholars.

Cath (2018), for instance, raises the question on whether, when we take the industry's perspective on governing AI through the lens of fairness, accountability and transparency, we are not leaving a considerable amount of issues unanswered and pushing forward values that are specific to the USA, where a large amount of companies working on AI are based, to the detriment of values particular to the Global South, for instance. Further, Mittelstadt (2019) argues that there is a huge gap in putting principles into practice that makes ethic-based frameworks weak strategies to stop machine learning pitfalls.

Commenting on these issues, Black and Murray (2019, p. 15) draw attention to how the libertarian, anti-regulation discourse particular to the beginning of the history of Internet is also to blame in the development of oligopolies by Big Tech companies, and argue that "[i]f we are to seek to control the way corporates and governments use AI and ML, then ethics cannot substitute for law or other forms of formal regulation". They thus stress that we need a robust, holistic and coherent system for regulating the development and use of these technologies that goes beyond transparency and also probably beyond sectoral regulations, which can give rise to "patchwork regulation in which there are overlaps and underlaps, with conflicting goals and logics".

Commenting regulation in general, and not strictly regarding AI, Baldwin and Cave (2021, p. 4) affirm that the debate over regulating or not is misleading. Law has been a fundamental propeller in allowing for the concentration of private property, the reliability of contracts and the development of capitalism as a whole. As the authors put it, "[l]aws allow markets to operate. We may object to examples of bad regulation, to controls that impose costs unjustifiably, but these examples make no case for seeing

regulation negatively". In this sense, it is our role to think of ways to create regulation that is effective in protecting rights of individuals, especially in face of issues that they do not understand but that strongly influence their lives. These ideas are to a great degree the same as those shared by Ayres and Braithwaite (1992).

Law proposals aiming to regulate AI have been under debate in different countries and regions such as China, Brazil, USA and the European Union, each with very different perspectives. They range from laws focusing on the regulation of specific systems, such as in China's Internet Information Service Algorithmic Recommendation Management Provisions, which rules content recommendation systems (CHINA, 2022); on impact assessments, such as in the US' Congress bill Algorithmic Accountability Act of 2022 (UNITED STATES OF AMERICA, 2022); on a risk-based approach that encompasses obligations to systems posing high or unacceptable risks to society, as is the case of the EU's AI Act (EUROPEAN UNION, 2020); and on a principles-based approach, as is the case of the first draft of the Brazilian Legal Framework for Artificial Intelligence (BRASIL 2020c), which might be considerably modified by a Commission of Jurists in Senate in the following months.

As mentioned above, it is not our goal here to assess these legislations in detail, but instead to understand how the Brazilian General Data Protection Law (LGPD) rules the development and deployment of ML systems with a specific focus on transparency. Data protection is insufficient to tackle all of the risks of ML, as it is incapable of addressing every single negative impact of machine learning, such as the environmental impacts[10] and the abuse of workers that are involved in the life cycle of some of these systems. It also does not address the impacts suffered by groups that are discriminated by these systems through the use of anonymised information, that do not fall within the LGPD's or GDPR's concept of personal data (TAYLOR, 2017).

However, as already highlighted in Chapter 4, data protection regimes do provide us with important mechanisms for the protection of rights of individuals, due to the sig-

---

[10] Some have argued that when reducing ads to comply with the GDPR, sites can significantly reduce energy expenditure and thus carbon footprint. See, e.g., ADAMS, Chris. How much CO2 can you save when you remove ad-tracking from news sites? **Chris Adams Blog,** 27 May 2018. Available at https://blog.chrisadams.me.uk/posts-output/2018-05-27-how-much-co2-can-you-save-when-you-remove-ad-tracking-from-news-sites/. Last accessed 14 July 2022.

nificancy that personal data plays in both the training of these systems and in the outputs that they issue. Therefore, it is now our role to understand how, based on the LGPD, can we stimulate and ensure the effective transparency of ML systems in order to achieve the goals we have outlined, such as broaden the understanding on how these systems shape society, increase accountability, among others.

At this stage, we aim to do so through an assessment of how the regulatory instruments provided by the LGPD can assist in the effective application of its transparency provisions. The broad meaning of these rules, that open the possibility of different ways of interpreting how to effectively comply with them, requires from regulators and other players further detailing on how compliance should be achieved. This could be made either *ex-ante*, such as through the publishing of guidelines, or *ex-post*, through the imposition of sanctions accompanied by the interpretation of these provisions.

I see a parallel in the LGPD with what Kaminsky and Malgieri affirm about the GDPR that it provides for a systemic, collaborative governance regime. In this framework, data controllers and regulators would jointly establish appropriate safeguards and transparency criteria for the processing of personal data, including by automated means, through ongoing conversations.

Such dialogues would take place through the "development of government guidelines and potentially involving industry-wide efforts to come up with codes of conduct or other forms of standards" (KAMINSKY, MALGIERI, 2019, p. 7; WP29, 2017). This logic is expressed, for instance, in Article 40, GDPR, which provides for the adoption of industry-wide codes of conduct and standards, and reflected in Article 50, LGPD, which has a very similar wording. As such, the systemic governance regime of these laws would be key to allow regulators and controllers to collaboratively assess the degree of information provision necessary for each system.

I add that such actions should include other players involved in the LGPD's regulatory apparatus, including the *Conselho Nacional de Proteção de Dados* (National Data Protection Council, CNPD), other regulators besides the National Data Protection Au-

thority (ANPD), standardisation bodies, civil society organisations, affected communities, among others, depending on the case.

Hence, in this Chapter, we will map how should LGPD's transparency rules be detailed and enforced regarding ML systems by the ANPD in an interplay with other actors based on the regulatory instruments that the LGPD provides.

Given the complexity of this task, this work will assess the suitability of the theory of responsive regulation to guide the enforcement of the LGPD with regard to ML systems. It will take into consideration not only whether its idea of flexibility if suitable to promote transparency in adapting to different systems and data controllers, but also because the ANPD has already signalled that its enforcement would be guided by this theory (URUPÁ, 2021; BRASIL, 2021b). As such, it is our goal to assess what benefits can we extract from this theory and what are the limitations it may have in the field we aim to assess, i.e., ML transparency, as well as in Brazil, a country from the so-called Global South (for the lack of a better term) which has a very different regulatory environment in comparison to Australia, where the theory was designed.

### 5.1. Responsive Regulation: First Thoughts

Based on our findings from the previous pages, the regulation of ML systems, especially for the enforcement of the framework of the LGPD with regard to transparency, should take into consideration two main features.

First, it should be capable of allowing for sufficient malleability of the regulator in order to implement policies that are suitable for addressing different ML systems in different contexts. As we mentioned previously, the kind of ML application and the context in which it is applied will lead to the necessity of assessing distinct information.

Second, it should allow for a constant dialogue not just between regulator and regulated entities, including their data protection officers (DPO), or *encarregados* under the LGPD, who are the individuals named by the controller and processor to act as a channel of communication between the controller, the subjects of such data and the

ANPD (Article 5, VIII, LGPD). It should also be extensive to other interested stakeholders including civil society organisations, researchers, auditors, representatives of affected groups, and many others. Each of these actors may play a different role in the oversight of data controllers regarding their compliance with transparency rules.

With that in mind, it is now our role to assess the suitability of the theory that the ANPD has signalled as the most suitable for its goals, the responsive regulation, to achieving an effective regulation of ML systems' transparency.

The responsive regulation theory came to light with the purpose of transcending the stalemate between those who advocate for more regulation and those who favour deregulation. By arguing that good regulatory policy is about "understanding private regulation — by industry associations, by firms, by peers, and by individual consciences — and how it is interdependent with state regulation", Ayres and Braithwaite proposed that, in most cases, the mix between public and private regulation opened up effective possibilities for addressing socio-economic issues arising in different markets (AYRES & BRAITHWAITE, 1992, p. 3).

The authors do not propose a clearly defined program or an ideal roadmap for regulators to apply. Instead, they highlight that "the best strategy is shown to depend on context, regulatory culture, and history. Responsiveness is rather an attitude that enables the blossoming of a wide variety of regulatory approaches" (AYRES & BRAITHWAITE, 1992, p. 5).

The theory aims thus to overcome the simplistic regulate-deregulate debate by claiming that, for a regulation to be effective, it should create rules that incentivise a regulated entity to voluntarily follow them, through strategies that range from granting awards and quality seals to unbearably harsh punishments (AYRES & BRAITHWAITE, 1992). All this should be made in an environment of constant dialogue between regulator and regulated, in a way that reduces information asymmetries between these actors and help better align the interests of regulated actors and society (ARANHA, 2019) by shaping the behaviours of actors in the regulatory game (AYRES & BRAITHWAITE, 1992, p. 44).

As Aranha (2019, p. 100) puts it, the responsive regulation theory has considerably evolved since the 1980s. Much work has been done, including by other theorists over the first works by Ian Ayres and John Braithwaite, who traced the grounds for the theory. Braithwaite (2010) himself says that the theory is a "collective creation". We will make reference to this further research as important complements to the authors' first works, especially considering that the context to which the theory was designed, reflecting peculiarities mostly of Anglo-Saxon countries, is quite different compared to the Brazilian particularities. Nevertheless, the theories main premises persist and are of great value for our analysis.

### *5.1.1. To regulate or to deregulate // to punish or to persuade: these ain't the questions*

One of the main theoretical assumptions of responsive regulation is the inability of law and procedure to simultaneously cope, only by themselves, with all of the goals they aim to achieve. Regulators are, according to Braithwaite (1984, p. 376), in a much stronger position when they have at their disposal the power not only to impose sanctions such as fines when a regulated entity is not in full compliance with the law. Instead, they are much stronger when they have at their disposal a menu of actions that range from bargaining power towards persuasion instead of punishment (BRAITHWAITE, 1985), the ability to make structural reforms and even to give prizes and awards to regulated agents when they have an exemplary conduct (KOLIEB, 2015).

In this sense, cooperation between regulators and regulated agents is fundamental to promote compliance through effective negotiation between businesses and state agencies. This dialogue leads to a better understanding of the market and also of greater trust between these actors, which enhances their capacity to cooperate.

Responsive regulation is, hence, mainly about finding the right balance between punishment and persuasion in order to make regulation effective. Under this rational, its theorists argue that, when regulators adopt strategies based strictly on punishment, their actions undermine the good will of actors when they are motivated by a sense of

responsibility. On the other hand, however, when the strategy is based totally on persuasion and self-regulation, state action will probably be exploited when actors are motivated exclusively by economic rationality (BRAITHWAITE, 1985).

### *5.1.2. Tit-for-tat*

Understanding that regulated entities have different motivations for complying with the law is crucial, and lead to the notion that corporate actors are "bundles of contradictory commitments to values about economic rationality, law abidingness and business responsibility" (AYRES & BRAITHWAITE, 1992, p. 19).

For that reason, each market and business will require that the state regulator take different approaches to guarantee compliance, as their motivations differ from one to the other. Some regulated agents will naturally be more willing to comply with the law, and thus persuasive approaches through negotiation might be more suitable for them than the imposition of harsh sanctions irrespective of their historic as well-intentioned actors for an occasional legal violation. For others that find law as a mere obstacle for economic gain and are constantly trying to evade it, perhaps persuasion will not be effective, and thus require harsher punishments.

In this sense, the authors support the adoption of a tit-for-tat (TFT) strategy, which consists of a mixture between punishment and persuasion that is both provokable and forgiving, and that they find more likely to be effective that choosing strictly one of another for any regulatory action (AYRES & BRAITHWAITE, 1992, p. 5). They aim to propose an alternative to regulatory strategies based strictly on command and control, here understood as the design of rules with concrete and preferably exhaustive commands for entities to avoid certain behaviour or to act in a certain way under the threat of a sanction in case of non-compliance (BALDWIN, CAVE, 2021, p. 30), a form of micromanagement of private activity. Instead, the responsive regulation is based on the creation of internal incentives through the alignment of interests of society and regulated entities through the state, regulatory authority (ARANHA, 2019, pp. 84-5).

According to the authors, it is thus necessary to overcome the apparent incompatibility between the ideas of those who think that regulated entities will comply with the law only when confronted with tough sanctions, and thus strictly based on the fear of being punished, and those who believe that gentle persuasion is enough to guarantee compliance (AYRES & BRAITHWAITE, 1992, p. 20).

Based on empirical work conducted interviewing executives, employees and government agents during the 1980s and early 1990s, mostly in the US and Australia, the authors argue that corporate actors are not only concerned with maximising profits or reputation, but they are also concerned with doing what is right, abiding to law and sustaining a self-concept of social responsibility (AYRES & BRAITHWAITE, 1992, p. 22). Other elements identified by the authors as motivations for corporations to comply with the law were

> (…) intangible consequences of adverse publicity for corporate prestige and employee morale, the harrowing experiences of senior executives in dealing with protracted cross-examination, and the dislocation of top management from their normal duties while they defended the corporation against public attack (BRAITHWAITE, 1985, p. 90).

Due to these various kinds of motivations held by executives, Ayres and Braithwaite conclude that a regulatory strategy that takes only punishment into consideration, and thus consider the figure of the regulator as merely a sort of avenger who dialogues with regulated entities only to impose sanctions, will fail to ensure compliance.

Instead, regulators should always try to be attentive to different businesses' and markets' characteristics, so as to identify their motivations and thus better define when and how to take action, and how to design norms and strategies that are most suitable to their different realities. For actors motivated strictly by an economic rationality, whereby a constant calculus of the pros and cons of not complying is present, punishing strategies will probably be more suitable. For actors motivated by a sense of responsibility, regulators should tend towards persuasion (AYRES & BRAITHWAITE, 1992, p. 19), which means finding ways to promote compliance by advice, education and entreaty (BRAITHWAITE, 1985, p. x), or, in general terms, through means other than imposing sanctions.

In this sense, regulatory objectives are, according to the authors, more easily achieved "when agencies display both a hierarchy of sanctions and a hierarchy of regulatory strategies of varying degrees of interventionism" (AYRES & BRAITHWAITE, 1992, p. 6). State intervention on businesses escalates and de-escalate in accordance to the level of compliance of the regulated entities. In this context, the heavier are the sanctions that regulators have at their disposal, that can even go beyond fines to allow them for a complete restructuring of the board of a regulated entity or to withdraw its licences to operate, more expressive is its capacity to ensure enforcement through persuasion. Or, in the words of the authors, "[p]aradoxically, the bigger and the more various are the sticks, the greater the success regulators will achieve by speaking softly" (AYRES & BRAITHWAITE, 1992, p. 19).

Excessive punitive measures may lead to legal resistance by regulated entities (ARANHA, 2019, p. 107) that can be expressed by an excess of judicialisation, a consequential lack of the payment of fines by offenders and thus less enforcement. This is what happened in Brazil, where a study of the Union's Court of Auditors concluded that, between 2011 and 2014, only 6.03% of the fines applied were actually paid (TRIBUNAL DE CONTAS DA UNIÃO, 2017; ARANHA, 2019, p. 57). It is important to notice, moreover, that this shows also the deficiency of strategies limited to deterrent sanctions such as fines. To deal with irrational actors, Ayres and Braithwaite (1992, p. 30) argue, incapacitive sanctions are needed, such as license or charter revocations that have the power to pull a company out of a market.

This leads us to the notion that the kind of regulatory strategies to be adopted by regulators that aim to follow the responsive regulation theory is — similarly to ML transparency — context-dependent. This means that the characteristics of the regulated agent, its economic and political power, the peculiarities of the technology, the market in which it is being applied, the consumers using it, of affected individuals, among others, have all the ability to influence the strategies adopted by the regulator.

That is why flexibility is crucial for the theory to be put into practice, as well as a thorough understanding of the market in question. After all, it is in the moments of interaction and of reciprocal influence between state and private regulation where the

authors of the theory consider lying the best opportunities to build the most suitable regulatory framework for each market and for each agent that would be an alternative to the regulate-deregulate debate (ARANHA, 2019, p. 103).

This includes even Braithwaite's (1985, p. 122) argument that self-regulation is not necessarily a softer option than public enforcement. Ideally, when a regulated entity has an effective and empowered compliance sector, it usually displays of more investigative and punitive capabilities, as well as the necessary information to prove offenders guilty, than an external party such as the government. As this most of the times is not the case, the author provides for a series of other regulatory strategies, as we will see further in this Chapter.

In any case, recent cases involving big corporations making use of ML systems show that such an idea of Braithwaite needs really to be thought about with a great degree of reflexion, especially when thinking of big digital corporations.

An example is the recent Facebook's whistleblower case, in which Frances Hughes leaked documents showing that the company knew its platforms had flaws that intensified the spread of disinformation and hate speech and did nothing about it even after its own employees raised red flags about the issue (WALL STREET JOURNAL, 2021). Another is Uber's recent leaks, that showed how the company's strategy to gain political and social support involved lobbying politicians through questionable means, breaking the law and even defending internally that violent protests should not be avoided because "violent guarantees success" (DAVIES et al., 2022).

These exemplary cases of regulatory entities' bad faith, however, do not necessarily debunk responsive regulation. The theory highlights the potential of self-regulation and other less intrusive regulatory mechanisms only in cases where regulators identify that regulated entities are virtuous (ARANHA, 2019). When their actions show otherwise and dialogue seems useless, "big sticks", as Ayres and Braithwaite (1992) call the strongest sanctions, should fall over their heads.

In order to see how this should be applied in practice in a regulatory enforcement framework, we now turn to an analysis to the theory's regulatory pyramids.

### *5.1.3. Pyramids ("êee faraó"[11])*

In understanding persuasion and punishment as interdependent and complementary in the design of regulatory strategies, the responsive regulation theory combines regulatory incentives and intrusive measures as means for the concretisation of regulatory objectives. As such, reflecting what was said above regarding the importance of an effective and powerful hierarchy of actions and sanctions for enforcement, another crucial aspect of the responsive regulation theory is the potential for escalation of regulatory intervention methods.

The application of this tactics takes place in the form of the so-called enforcement pyramids, which allow for regulators to escalate towards more intrusive constraints as offenders break more rules or act viciously. However, just as importantly, they also provide for the gradual de-escalation of constraints when regulated entities improve their compliance culture through time (AYRES, BRAITHWAITE, 1992, p. 35; ARANHA, 2019, p. 113).

The more compliant a corporate actor is, lower will be the degree of state intervention on its activities, thus inhabiting the lower levels of the pyramids. Nevertheless, in cases in which an entity starts to violate legal rules, the regulator will be legitimised to take action, imposing stronger sanctions and escalating the level of intervention, moving upward the pyramid as the offender moves towards a less compliant posture (AYRES, BRAITHWAITE, 1992, p. 6).

These pyramids should contemplate a hierarchy of sanctions and of regulatory strategies with varying degrees of interventionism. At the bottom of the pyramid lie the ones with a less degree of state intervention, while on the top should dwell the most threatening punishments, to be triggered far less often.

---

[11] GOMES, 2006.

It is important to note that the pyramids should preferably not contain only formal penalties, but also general ways of embarassing the regulated entities, such as an increase in frequency of inspections or the inclusion of oversight by citizens bodies (ARANHA, 2019, pp. 113-15).

Braithwaite (2006, p. 886) argues that it is an important presumption that the regulator should always try to start at the base of the pyramid with an offender, and then only escalate to more punitive actions when dialogue and modest forms of sanctions fail. That would apply, according to the author, even with the most serious matters, such as infringements of legal obligations by nuclear power plants operators, even if they are posing at risk thousands of lives. Nevertheless, the theory is sufficiently flexible to make any regulatory strategy possible, even by starting with the strongest sanctions if necessary (AYRES, BRAITHWAITE, 1992, p. 30). Such a formulation is interestingly made by Baldwin and Cave (2021), as we will discuss later.

For the theorists, the greater the level of enforcement to which regulators can escalate in the pyramid, the greater will be the willingness of regulatees to comply. In other words, “[r]egulatory agencies will be able to speak more softly when they are perceived as carrying big sticks" (AYRES, BRAITHWAITE, 1992, p. 6).

This approach is the essence of the aforementioned tit-for-tat strategy, whereby “the regulator refrains from a deterrent response as long as the firm is cooperating; but when the firm yields to the temptation to exploit the cooperative posture of the regulator and cheats on compliance, then the regulator shifts from a cooperative to a deterrent response” (AYRES, BRAITHWAITE, 1992, p. 21). Nevertheless, it is fundamental to understand that, as businesses’ and markets’ motivations and dynamics vary to one another, different levels of enforcement and strategies will have to be designed by the regulator, almost in a tailored manner for each corporate actor.

Two examples of regulatory pyramids presented by Ayres and Braithwaite are of particular importance for us. The first is the enforcement pyramid and the other a pyramid of regulatory strategies. We will start by analysing the enforcement pyramid.

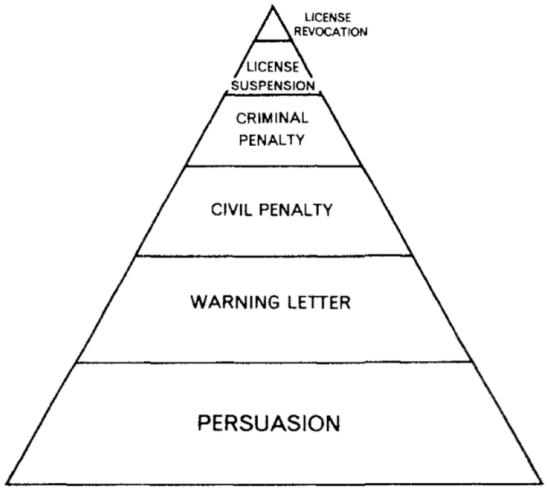


Figure 1: Example of enforcement pyramid
(AYRES, BRAITHWAITE, 1992, p. 35)

The enforcement pyramid provides for a range of different actions to be taken by the regulator according to the compliance of a regulated entity with the law. Ideally, most regulatory work should take place at the base of the pyramid, where regulators would be dedicated to make compliance possible only through persuasive methods such as dialogue and educative measures. In case persuasion fails to ensure compliance, the next step would be to send warning letters, then civil penalties such as fines. In case none of that succeeds in making an entity comply, the regulator would be in a position to apply criminal penalties, suspend licenses and, as a last resort, revoke the license of an entity to conduct a certain activity (AYRES, BRAITHWAITE, 1992, p. 36).

Having a broad range of applicable sanctions, including light and very harsh ones, is crucial for a strategy based on responsive regulation to be successful. They should not be limited to the ones expressed in the example proposed by the authors, but can also consist of measures for restructuring boards of executives, for instance (BRAITHWAITE, 1985). The pyramid intends to provide regulators with a strong bargaining power against regulated agents, since there will always be the risk for the latter to be subject to serious penalties in case of non-compliance.

The second pyramid proposed encompasses regulatory strategies, a term that reflects the integration of different regulatory instruments with the aim of influencing social

behaviour. Instruments (also called techniques), on their turn, are the means deployed by the state to influence such behaviour and thus achieve the goals underlined by public policies (ARANHA, 2019, p. 68). These instruments can be, for instance, legal commands; wealth deployment, such as contracts and subsidies and the disclosure of information, all of them for the purposes of influencing behaviour (BALDWIN, CAVE, LODGE, 2012, p. 106).

Within the strategies pyramid, the authors, as we mentioned above, consider that, whenever possible, self-regulation can be a good option for ensuring compliance due to the capacities of industry to understand its own mechanisms. Once an agent is not compliant under a self-regulatory framework, the regulator would thus gradually escalate in the pyramid. If the last level is reached, towards what Ayres and Braithwaite (1992, p. 38-9) call "burning of bridges", i.e., the regulators would be able to adopt a strategy based on command regulation with nondiscretionarity.

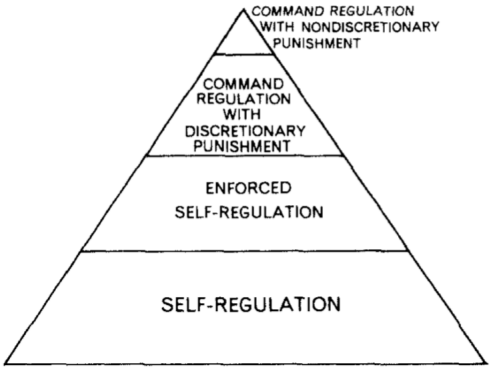


Figure 2 - Example of a pyramid of enforcement strategies
(AYRES, BRAITHWAITE, 1992, p. 39)

Once again, the authors affirm that this pyramid is just an example, and that any framework of enforcement strategies should be designed by a regulator taking into consideration the context in which it is acting to enforce legal obligations (AYRES, BRAITHWAITE, 1992, p. 38). Moreover, they argue that an important step, aligned with the idea that self regulation can be very effective in ensuring compliance, is to recognise the important role that industry associations have in advising individual

firms to cooperate. Involving these associations in the oversight of the market is thus a valid tool to be included in a regulatory strategy (AYRES, BRAITHWAITE, 1992, p. 39).

### *5.1.4. Responsive Regulation and the Global South: thinking beyond centralisation*

In his article "Responsive Regulation and Developing Economies", John Braithwaite (2006) acknowledges that there are particular characteristics in countries from the Global South to apply responsive regulation, an approach designed in and mostly for the Global North (AYRES, BRAITHWAITE, 1992). He thus draws attention to what he refers as a smaller regulatory capacity in the Global South and to a supposedly minor oversight by NGOs and social movements to mobilise as potential inhibitors of regulatory success (BRAITHWAITE, 2006, p. 885).

One may question whether his assumptions are based on empirical work since he frames this particular work on the basis of critiques made by authors from the Global North, as Braithwaite (2006, p. 884) himself argues. In any case, some of the aspects he brings are of interest for this work, as the differences highlighted by the author between these, let's say, two different regions of the world are worthy of consideration and result from extremely particular issues that range from cultural features to budget, legal traditions and so on.

To illustrate this, when the author affirms that "[d]eveloping countries mostly have less oversight by [Non-Governmental Organisations -] NGOs and social movements to mobilize" (BRAITHWAITE, 2006, p. 885), he does not make reference to any particular empirical study. Nevertheless, the divide between NGOs in the Global North and the Global South can be identified, at least indirectly, when we assess the differences in terms of both funding and representation in institutional fora, where Global North NGOs are much more represented especially due to their larger budget than their counterparts from the Global South (SÉNIT, BIERMANN, 2021; GEREKE, BRÜHL 2019). Although such a figure does not directly reflect NGOs' capacity to mobilise at a national level, it may be an influencing factor to understand the ability

of NGOs to push forward their agendas, as issues such as the quantity and expertise of employments are highly influenced by budget.

A brief note is worth making, nevertheless, with regard to civil society engagement in Brazil concerning data protection and other digital technologies and internet governance issues. Most non-for-profit organisations and think tanks working with these topics in Brazil are now part of a coalition of more than 50 members, the *Coalizão Direitos na Rede* (Rights in the Network Coalition - CDR)[12], where activists and researchers have a room to debate and articulate for public action within the scope of their activities.

Another difference highlighted by Braithwaite (2006) between Global North and South governments with regard to regulatory strength relates to their enforcement capacities. Alone in the field of data protection authorities, a good example of the different capacities between regulators in different jurisdictions is the fact that the Brazilian authority, ANPD, has currently 71 staff members[13], while the German authority, the *Bundesbeauftragte für Datenschutz und Informationsfreiheit* - BfDI, had 270 employees as of 7 January 2019 (BfDI, 2019), even though it is competent for supervising only Federal public bodies and commercial providers of telecommunication services (GERMANY, 2017). ANPD, on its turn, is responsible for the oversight of every data controller or processor in Brazil falling under the scope of the LGPD.

Moving forward, an important aspect of Braithwaite's (2006) article is his description of the evolution of responsive regulation, which led in his view to a system of accountability that is much more deliberative, circular and democratic, in which the state no longer holds a position of ultimate guardian of regulatory rules.

Instead, there would be multiple regulatory guardians, including NGOs, audit offices, ombudsmen, courts, public service commissions, self-regulatory organisations, among others, responsible for the oversight of legal rules and that would hold not only regu-

12 https://direitosnarede.org.br/

13 Information obtained from a staff member at the ANPD during interview.

lated entities, but also everyone else in this circle accountable for their actions, be it a state or non-state actor. (BRAITHWAITE, 2006)

This is what Braithwaite (2006, p. 886) would call “nodes of networked governance”, and each node would have a role not only in ensuring compliance but also to check abuse of power by other nodes and also whether they are sufficiently autonomous or not to exercise their respective roles.

Braithwaite thus draws attention for the central importance of NGOs in the regulatory oversight, and that they would also assume a rise if directly regulating businesses through, for instance, “naming and shaming, restorative justice, consumer boycotts, strikes, and litigation” (BRAITHWAITE, 2006, p. 888). In this sense, one alternative for regulators with less capacities is that they, instead of escalating in terms of state intervention, escalate in terms of state networking with non-state regulators, by gradually including more non-state parties in the oversight of a market or entity (BRAITHWAITE, 2006, p. 890) as expressed in the following pyramid:

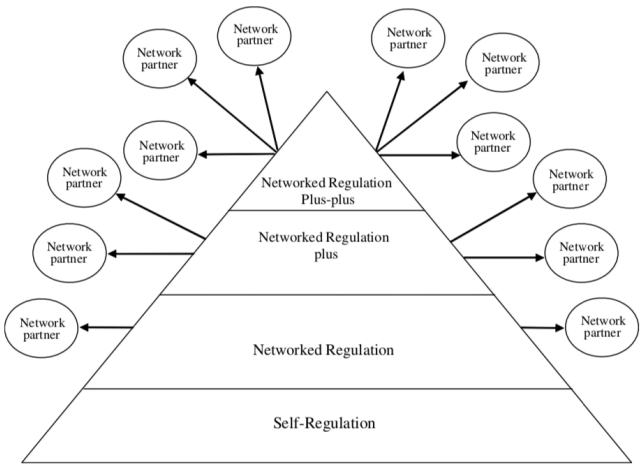


Figure 3 - Example of a pyramid for networked governance
(BRAITHWAITE, 2006, p. 890)

The pyramid expresses how enforcement can range, when regarding the involvement of other actors beyond the regulator to conduct oversight, from the inclusion of more

actors in inspecting the work of regulated entities, from strategies involving self-regulation until networked regulation with many oversight players.

#### *5.1.5. A Collective Creation*

As we mentioned above, in the years that followed the publishing of Ayres and Braithwaite's seminal book from 1992, other features were added to the development of the theory.

Among them, we could highlight (i) the regulatory diamond proposed by Kolieb for including in the theory's rationale mechanisms to reward regulated agents for adopting measures that go beyond the mere compliance with law (KOLIEB, 2015); and (ii) the idea of networked governance, which relates to the creation of a "regulatory society" where NGOs, audit bodies and local social pressure would play a key role in regulatory efforts, especially in developing countries (BRAITHWAITE, 2006). Such a participatory framework allows for diverse groups of interest to take part and provide inputs in the regulatory dynamics, ensuring that regulators and regulated actors address more properly the needs of different social groups.

One theory that has similar features as ones of the responsive theory and that has interesting lessons for us to in the field of regulatory enforcement is the *positive regulation theory*, described by Baldwin and Cave (2021). We will address its main ideas in the next section as a way to widen our understanding of possible regulatory strategies that can be used in the enforcement of the LGPD.

### 5.2. Contributions to Regulatory Enforcement from the Positive Regulation Theory

The positive regulation theory aims to provide a framework that is at the same time pro-business and pro-society. It has also a strong focus on the idea that the optimal regulatory scenario is the one in which regulation is effective in harnessing the self-regulatory capacities. This is seen by the authors as the most cost-efficient way of

regulating, as it allows the state to transfer regulatory costs to the industry by steering corporate power in productive and useful directions (BALDWIN, CAVE, 2021).

As such, it differentiates itself from responsive regulation as the latter is not necessarily aimed at fostering businesses activity, but instead to make regulation the most effective possible in achieving its goals. This goal is more aligned with the general framework of data protection regimes, that aim to protect a fundamental right. In any case, positive regulation has important contributions that we should take into account.

Quite similarly to Ayres and Braithwaite (1992), Baldwin and Cave (2021, p. 69) affirm that the success of this strategy depend on a myriad of factors such as the organisational culture of the agents in the industry involved, the complexity of risks the market poses, the rate of change in the sector, among others. As such, they argue that, in most regulatory contexts, regulatory strategies and intervention styles will have to be mixed in order to adapt to the specificities of each activity and agent (BALDWIN, CAVE, 2021, p. 68).

Under positive regulation, enforcement is based on five steps that consolidate what the authors call the "DREAM framework". The first is *detection*, which consists of "the gaining of information on non-compliant and undesirable behaviour". The second is *response development*, related to "the developing of policies, rules, and tools to deal with the problems discovered". The third is *enforcement*, which is the "application of policies, rules, and tools on the ground", followed by *assessment,* "the measuring of success or failure in response development and enforcement activities". The last step is *modification*, which is the adjustment of "strategies and tools in order to improve compliance and address problematic behaviour", which allows for a frequent self-analysis by the regulator to understand to what extent the actions taken are being effective to achieve its goals (BALDWIN, CAVE, 2021, p. 70).

Similarly to what happens in responsive regulation, positive regulation considers that there are strengths and weaknesses in what they call compliance and deterrence strategies (BALDWIN, CAVE, 2021, p. 84). Compliance strategies are similar to what Braithwaite (1985) calls persuasion, and consists mostly of negotiations and ed-

ucative actions. Deterrence, on the other hand, englobes the ideas which Braithwaite (1985) include under the punishment umbrella.

At this point, Baldwin and Cave (2021, p. 89) present a critical analysis of the idea of the pyramids and of its gradual escalation. The first difficulty they pose relates to the fact that, where risks of catastrophic outcomes are in place, or when such outcomes have already occurred, it might not be wise to enforce law by escalating step by step the layers of the pyramid. Instead, immediate action at an upper layer such as, for instance, forcing the agent to stop its activity, might be necessary.

That would be the case, for instance, of a well-resourced credit bureau company — and very capable of hiring good compliance experts — which, after the approval of the LGPD, sold datasets with information about indebted people to marketing agencies without their knowledge. As this would probably be a violation of the law's principles of purpose limitation, necessity, adequacy and transparency, to say the least, it would be hard for the Authority to explain that education activities would be justifiable.

For the authors, an additional hindrance in escalating the pyramid is that it may simply not happen due to resource constraints that may impede it, the fear of political consequences and the lack of information about potential improvements in the regulated agent compliance (BALDWIN, CAVE, 2021, p. 90).

Additionally, moving down the pyramid to decrease the punitiveness of the approach may also not work every time. There are cases in which the application of punitive sanctions may affect trust or good will between regulators and regulated agent. Without a good relationship, acting persuasively would fall outside the scope of responsive regulation (BALDWIN, CAVE, 2021, p. 90).

Other challenges for the escalation of the pyramid rise in markets where regulators with different strategies, mindsets and missions can have a say in the enforcement of law (BALDWIN, CAVE, 2021, p. 90). Data protection is full of such examples, including when considering the adoption of machine learning, as data controllers can be under the authority of many agencies at the same time. Hospitals, for instance, may have to comply with data protection rules enforced by ANPD and with health-related

rules enforced by the *Agência Nacional de Saúde Suplementar* (National Agency for Supplementary Health - ANS). As each of these agencies have different strategies, being able to make them work side by side in the enforcement of the layers of the pyramid might not be as straightforward as it might seem at first sight.

### *5.2.1. The roles of risk, capacity and good intentions*

In this sense, in order to present their own proposal for regulatory enforcement, Baldwin and Cave (2021) make use of a combination between responsive regulation and risk-based approaches. Having traced some limitations of the former theory, they also affirm that the latter, despite having expressive strengths, should not have its weaknesses disregarded.

Risk-based frameworks are noteworthy for allowing regulators to determine priorities to tackle the activities that are supposedly most problematic for society (BALDWIN, CAVE, 2021, pp. 76-7).

Nevertheless, as the process of determining risk degrees is far from neutral and involve a complex set of choices, biases and evaluations, frequently the defined risks are not future-proof or able to answer properly to reality. Instead, such definition of risks depend on factors such as how regulators or legislators are being informed and how political and industry pressure, as well as public opinion, are shaping their view (BALDWIN, CAVE, 2021, pp. 77-79).

Further, the process of prioritising risks make clear what should prioritised — but also what should be ignored. In this sense, risks that are not identified at first hand by the regulator will fall out of its radar, giving rise to potential burdens which may be left unaccountable, as, in theory, the damage creator was not in an obligation to cope with the same rules as actors under a higher risk category.

As such, a risk-based regime has the potential of attaching the regulator to a certain basket of risks that might become easily out-dated (BALDWIN, CAVE, 2021, p. 80) and that “gives no indication of the extent to which undesirable risk creation is escaping the regulatory net” (BALDWIN, CAVE, 2021, p. 81). In this sense, definitions of

risk should be regarded more as a “way to construct the regulatory agenda rather than a mechanical solution to the familiar challenges of regulation” (BALDWIN, CAVE, 2021, p. 81), and thus be, although not the basis for the regulation, a complementary part of it, to be determined in a flexible way through time.

That said, Baldwin and Cave (2021) propose a method for determining regulatory strategies that they call Good Regulatory Intervention Design - GRID. Within its framework, a regulator should take into consideration, when determining the degree of intervention in a regulated entity, three main elements.

The first regards the costs involved in the enforcement of different enforcement tools, that should be graded as low, medium or high cost. Although affirming that costs may vary according to context, an example of low cost tool would be a determination to regulated entities to self-monitor and self-certify. The LGPD provides for such a possibility as a means for a data controller to be able, for instance, to transfer data outside the Brazilian borders (Article 33, II, d, LGPD); or the adoption of good practices being an element for potential penalty reductions (Article 52, §1, IX, LGPD).

An example of medium cost tool would be, on its turn, conducting random monitoring of regulated entities or audits of control systems. Under the LGPD, the ANPD would be allowed to conduct auditing in ML systems according to Article 20, §2, which we have debated in the last Chapter, about the explainability of automated decision-making systems.

On its turn, an example of high cost tool is the conduction of inspections on site. This might also be allowed under the LGPD under the same Article 20, §2, or under Article 55-J, XVI, which allows the Authority to audit data controllers and processors under the scope of an investigation (BALDWIN, CAVE, 2021, p. 97).

The second factor for influencing a regulatory strategy brought by the authors is the type of regulated entity, according to their intentions and capacity to comply. They classify firms as (BALDWIN, CAVE, 2021, p. 98):

1. Well-intentioned and with high capacity to comply;

2. Well-intentioned and with low capacity to comply;

3. Ill-intentioned and with high capacity to comply;

4. Ill-intentioned and with low capacity to comply.

For the authors, intervention would be stronger as they progress down this list. This means that the more ill-intentioned, and lower the capacity to comply, the more intrusive and heavy would be the interventions. Regulators should, similarly to what happens in responsive regulation, be attentive to changes in the natures of regulated entities and act accordingly by strengthening or weakening the interventions (BALDWIN, CAVE, 2021, p. 98).

Finally, the last factor is the type of risk. The higher the risk, higher should be the intervention. Regulators should, however, always have in mind, as we said some paragraphs above, that risks may change over time and thus flexibility and responsiveness is crucial to address this. The authors thus propose five levels of risk (BALDWIN, CAVE, 2021, p. 99):

| | |
|---|---|
| **Low risks, stable** | Inherent low risks or net low risks, the levels of which are not likely to change in the periods between regulators' risk reviews. |
| **Low risks, unstable** | Inherent low risks or net low risks the levels of which may change in the periods between regulators' risk reviews. |
| **Medium risks, stable** | Inherent medium risks or net medium risks the levels of which are not likely to change in the periods between regulators' risk reviews. |
| **Medium risks, unstable** | Inherent medium risks or net medium risks the levels of which may change in the periods between regulators' risk reviews. |
| **High risks** | High risks that are likely to remain so. |

Table 3 - Risk levels under the GRID

(BALDWIN, CAVE, 2021, p. 99)

Within the table, “inherent” risk is used when the risk of an activity or product cannot be reduced through good risk management, while “net” risk is one that can be reduced through good governance.

The criteria of stability/instability, on the other hand, refers to whether the risks are likely or not to change over time. As the authors put it, shifts in the nature of an activity, such as when, for instance, chemical materials are changed in a mining activity, can lead to changes to inherent risks. Net risks, differently, may change when there are modifications in the quality of the risk management system used by a regulated entity (BALDWIN, CAVE, 2021, p. 99).

I would add that changes in risk can also happen when the perception about a specific activity or product among society changes. It is the case, when regarding machine learning, of its use within social media applications. Through time, with the spread of disinformation and hate speech in platforms like Twitter, Instagram and TikTok through the use of algorithms and the effects these had in democratic processes around the world led to a new perception of the impact of these machine learning systems and the environment, context in which they were built (Big Tech companies). Society perception  about social media changed in the meantime, from naïve tools for making friends to potential shapers of democratic debate (HOWARD, 2020).

### *5.2.2. The GRID*

With that being said, Baldwin and Cave (2021) present a figure with a gradient that represents how regulatory intrusiveness and costs may increase according to the type of regulated entity and its ratio based on intention/compliance capacity, and the type of risk, whether high or low, stable or unstable. The arrows represent the increase in intrusiveness:

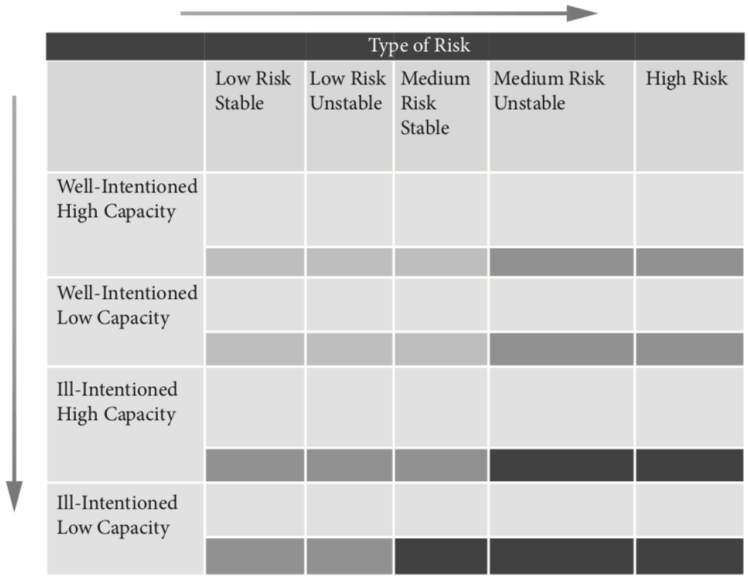


Figure 4 - The GRID (Good Regulatory Intervention Design), as proposed by Baldwin and Cave (2021, p. 100)

The GRID proposed by Baldwin and Cave (2021) does not need to be seen as exclusionary of the pyramid framework proposed under the scope of responsive regulation. Instead, it makes more tangible the assessment of regulated entities when designing regulatory strategies suitable for them by making more concrete the factors to be assessed in this process.

At the same time, the GRID leaves the door open not only to include other factors in the assessment, but also to not take capacity or risks into consideration when they are not reasonable indicators of good or bad will according to the context.

For instance, in many cases involving the regulation of ML systems, the financial or workforce capacity of a small-sized startup might be irrelevant when considering the potential impacts that it might have with a data processing activity. With the democratisation in the access to automated data processing technologies and the easy access to large amounts of personal information, the activities of small and medium companies can present unforeseeable impacts to the right to data protection. In such situations, regulators can disregard capacity criteria in order to impose stronger sanctions.

As we will show in the next section, the ANPD has a range of different penalties at its disposal to enforce the law, including auditing powers that can also be useful if we expect to apply the GRID in the field of data protection in Brazil.

Moreover, although there is no particular provision allowing for the ANPD to apply sanctions based on the intention of the regulated entity, the capacity it has to comply is directly addressed by the LGPD. In its Article 55-J, XVIII, the law is clear that the Authority has the duty to provide for more flexible rules, proceedings and procedural time limits for small-sized companies and startups to comply.

## 5.3. LGPD, ANPD and Responsive Regulation

### *5.3.1. Enforcement*

Some of the provisions that allow us to argue that the LGPD, at least partially, meets with the responsive regulation's pyramids logic lies within its Article 52. It provides for a description of different penalties that can be applied by the Authority, that range from warnings and fines up to the prohibition of data processing activities. Within this framework, the law allows for the possibility of an escalation of sanctions, and thus for a pyramidal approach, considering that such sanctions can be applied on a "gradual, individual or cumulative basis" (Article 52, §1).

Also in its article 52, the LGPD establishes what are the criteria that the Authority needs to follow when applying sanctions. They include, among others, the good will of the offender and his or her level of cooperation with the investigations (Article 52, §1, II and VII); whether the offender is recidivist (Article 52, §1, V); whether the offender can prove the adoption of internal mechanisms and procedures capable of minimising the damage during the personal data processing (Article 52, §1, VIII); the adoption of good practices (Article 52, IX); and the prompt adoption of mitigative measures after the infringement was identified. There is thus legal room for the escalation of regulatory intervention to be dependant on the behaviour of the regulated and the results obtained (GARCIA, 2020, p. 55).

The LGPD's lighter penalties would be warnings (Article 51, I); fines of up to two percent (2%) of a private legal entity's, group or conglomerate revenues in Brazil, up to a total maximum of fifty million reais (R$ 50,000,000.00) per infraction (Article 51, II); daily fines subject to the total maximum referred to in previous penalty (Article 51, III); and the disclosure and publicisation of the infraction (Article 51, IV).

It is worth noting that the harshest sanctions in the LGPD, which consist of measures for the incapacitation of a service, are not supposed to be applied at first sight. These include sanctions for partial suspension of the use of a database related to a given infraction for 6 months (Article 52, X), for suspension of the personal data processing activity related to the infraction also for 6 months (Article 52, XI) and the partial or total prohibition of activities related to data processing (Article 52, XII). These penalties can only be imposed after other, less harmful penalties had already been imposed (Article 52, §6), something representative, by itself, of an escalation.

Dosimetry rules may provide for legal support in case the ANPD aims to push forward a responsive agenda with regulated entities. They allow for the Authority to take into consideration compliance efforts of data controllers and processors when deciding when and how to impose sanctions, with the harsher ones triggered by recidivism. Moreover, at least to what relates to the range of sanctions at the Authority's disposal, the freedom of not having to always have a gradual, escalating approach to each sanction provides ANPD with many possibilities to strengthen compliance with the LGPD. That, of course, with the exception of the heaviest penalties described in Article 51, X; XI; and XII.

Further, one should highlight that the principle established by the legislation in its Article 6, X, establishes that the data processing agent should demonstrate, before ANPD, "the adoption of effective measures capable of proving the observance and compliance with the rules of protection of personal data and, even, of the effectiveness of these measures". This notion is closely linked to the theory's approach for the enforced self-regulation strategy, which consists of a demand from the regulator to the regulated entity to internalise inspection costs through the creation of a department or group of internal compliance in order to monitor compliance with the rules and rec-

ommend disciplinary actions determined by the regulator against the offenders (ARANHA, 2019).

In October 2021, the ANPD published its first resolution, *Resolução CD/ANPD no. 1/2021* (also referred herein under as the "Resolution")[14], establishing rules for enforcing the LGPD, which provides crucial information on the regulatory strategies to be adopted by the ANPD in its enforcement.

The first aspect worth noting at the Resolution lies in its Article 15, where the Authority describes that its enforcement will be made through monitoring (*monitoramento)*, guidance (*orientação*), prevention (*prevenção*) and repression (*repressão*). The presence of these four different approaches for the enforcement of the LGPD reveal how the ANPD expects to push forward its regulatory strategy.

*Monitoring activities* are defined as the ones aimed at are gathering relevant information and data to support the ANPD's decision making with the purpose of ensuring the sound functioning of the regulated environment (Article 15, §1, *Resolução* CD/ANPD no. 1/2021).

Their objectives are fivefold: to plan and subsidise inspection activities with relevant information; to analyse the compliance of the regulated agents; to consider the regulatory risk in function of the behaviour of the regulated agents, in order to allocate resources and adopt actions compatible with the risk; to prevent irregular practices and foster a culture of personal data protection; and to support the regulated agent in repairing irregular practices and damages or, at least, minimising these damages (Article 18, I; II; III; IV and V).

These provisions show intents for regulatory responsiveness from the ANPD in terms of, for instance, an assessment of risk based not only in an abstract understanding of the activity, but also in the behaviour of the regulated agent. Moreover, activities such as the prevention of irregular practices and fostering of a data protection culture can be put into practice through a direct dialogue with data processors and controllers,

[14] All of the transcriptions in English for the rules established in the norm are made in free translation by the author.

even through the issuing of recommendations that make clear that, in case any chance of violation of law persists, harsh penalties can be imposed.

Another approach that might constitute lower layers in a regulatory pyramid is also seen under the *guidance* strategy provided by the Resolution. It is characterised by methods and tools aimed at guiding and educating data controllers, processors and data subjects (Article 15, §3, *Resolução* CD/ANPD no. 1/2021).

Guidance activities will consist of the publishing of good practice guidelines; recommendations for regulated agents to pursue further training (something quite odd); elaboration of self-assessment tools to be used by data controllers and processors; certification and incentivising the adoption of good practices and governance standards; among others.

The guidance approach may thus allow for meaningful dialogues between the ANPD and regulated entities in a way that allows both parties to learn from each other and for the Authority to guide the agent in optimising compliance. Self-regulatory tools are incentivised at this layer, and provide a set of tools suitable to more virtuous actors.

The *preventive* activities maintain this dialogical ideal, on their turn, through actions based, preferably, on the dialogue with the regulated agents who have violated the LGPD in order to find solutions that bring the agent back into full compliance. They also aim to avoid or remedy situations that could entail risk or damage to personal data subjects and other processing agents (Article 15, §3, *Resolução* CD/ANPD no. 1/2021).

The Resolution describes four actions under this approach (Article 32, *Resolução* CD/ANPD no. 1/2021). The first is the publishing by the ANPD of aggregate sectoral information and performance data on its website about a market or entity, such as the rate of problem resolution capacity and how many owner requests were fulfilled (Article 33, *Resolução* CD/ANPD no. 1/2021). The second is the issuing of notices by the ANPD to the regulated entity warning of potential violations to the law and information for the handling agent to identify the necessary steps to be taken in order to guarantee compliance (Article 34, *Resolução* CD/ANPD no. 1/2021). The third action

consist of requests for regulated entities to fine-tune their compliance (Article 35, *Resolução* CD/ANPD no. 1/2021). The fourth is the ordering for an agent to provide for a conformity plan establishing in detail the actions that the agent will take to reverse misconducts (Article 36, *Resolução* CD/ANPD no. 1/2021).

These regulatory tools may thus allow for the regulator to ensure compliance and to guide the regulated agent to improve its activities in order to prevent further violations in a way and promote optimised conformity with the law.

It is important to notice that there is no detail about which infractions will be dealt with through preventive activity, such as whether they are of lesser offensive potential or not. At the same time, the Resolution does not state that prevention will necessarily be the first strategy to be used by the ANPD when there is a violation.

On the one hand, this lack of detail can be positive, as it guarantees greater flexibility to the regulator, something foreseen in responsive regulation. On the other, negative impacts may arise if this freedom translates into arbitrariness in the sense that the regulator does not provide for sufficiently transparent parameters about how these regulatory tools will be applied. This may reflect in leaving both regulated entities and society in general in a context of legal uncertainty which may affect the regulatory strategy and compliance.

Finally, in the *repressive* framework lies ANPD's coercion tools, aimed at interrupting situations of damage or risk, bringing agents back into full compliance and punishing those responsible for such damages by applying the sanctions provided for in Article 52 of the LGPD by means of a sanctioning administrative process (Article 15, *Resolução* CD/ANPD no. 1/2021). These tools would thus probably dwell in the upper layers of the pyramid, even though, at least at the point of writing this dissertation, no clear idea of how the ANPD will make use of them, as no sanctions have yet been imposed by the Authority.

Making reference to the visual artefacts we saw previously in this Chapter, not necessarily should the monitoring, guidance and prevention approaches be seen as pertaining to different layers of a regulatory pyramid of a GRID gradient. They may, e.g., be part of a same layer with incentives for self-assessment in the basis of a pyramid. Ex-

amples are the fostering of a data protection culture among the monitoring activities under the monitoring approach (Article 18, IV, *Resolução* CD/ANPD no. 1/2021); the preparation and sharing of good practice guides and document models to be used by data controllers and processors under the guidance framework in the guidance approach (Article 29, I, *Resolução* CD/ANPD no. 1/2021); or the issuing of notices by the regulator under the prevention approach (Article 34, *Resolução* CD/ANPD no. 1/2021).

#### *5.3.2. Risk-based approach?*

Both the LGPD and the GDPR are frequently referred to as having introduced a risk-based approach for enforcing its obligations. This logic is seen by Doneda (2021) as a response to the spread of automatic data processing activities that over time led to an understanding that processing personal data would be a risk in itself.

It is worth assessing both laws once again side by side to understand their parallels and how scholars from both jurisdictions interpret their provisions to understand to what extent the general scope of these laws is indeed influenced by the notion of risk.

In the LGPD, one of the key provisions about risk and the role of its assessment in the enforcement of the legislation is its Article 44, II, which establishes that

> Article 44. Processing of personal data shall be deemed irregular when it does not obey the legislation or when it does not provide the security that its data subject can expect, considering the relevant circumstances of the processing, among which are:
> (…)
> II – the result and the risks that one can reasonably expect of it.[15]

Under the GDPR, a similar approach is found under its privacy by design and by default obligations, which calls data controllers to provide for the necessary safeguards

---

15 Original wording: "*Art. 44. O tratamento de dados pessoais será irregular quando deixar de observar a legislação ou quando não fornecer a segurança que o titular dele pode esperar, consideradas as circunstâncias relevantes, entre as quais:*
*(...)*
*II - o resultado e os riscos que razoavelmente dele se esperam*".

into the data processing in accordance with, among other variables, the risks involved in the operation (Article 25(1), GDPR).

As Menke and Goulart (2021) point out, risk assessments are an underpinning of information safety, and are directly linked to the principle of prevention, established by the LGPD in its Article 6, VII. As such, they are an inherent part of the compliance program of a data controller, since the category of risk of a data processing will determine what are the necessary safety mechanisms that will have to be put into place and, in some cases, even determine whether a data processing activity should be conducted at all.

They argue that, by focusing on risk, data protection regimes allow for the construction of a space of mutual trust between controllers and the authority, whereby controllers are given a vote of confidence for defining which of the activities they carry out amount for a larger risk, thus being obliged to provide for more substantial means to protect personal data (MENKE, GOULART, 2021).

The assessment of risk in the data processing and the security safeguards that are put in place by the controller for the enhancing protection, which includes the adoption of privacy by design measures, are also of great importance since they are one of the aspects to be assessed by authorities in case of data breaches or any other violation to data protection rights (MENKE, GOULART, 2021).

This is provided by Article 48, §3º, LGPD, which establishes that "[w]hen judging the severity of the incident, there will be an analysis [by the ANPD] of eventual evidence that, within the scope and the technical limits of the services, adequate technical measures were adopted to render the affected personal data unintelligible to third parties who were not authorised to access them." Similarly, GDPR establishes that, when assessing the imposition of fines for infringements, security and privacy by design measures shall be taken into consideration by the supervisory authority (Article 83(2)(d), GDPR).

It is thus appropriate to conclude that different degrees of compliance with the legislation are required depending on the risks involved in the data processing. When concerning automated decision-making systems, we may see transparency measures as

one of the measures to be implemented by the controller to enhance protection, especially when the system is responsible for making high-stakes decisions. As Gonçalves posits, "with the advent of big data, it is often not the collection of information in itself that is sensitive, but the inherently obscured inferences that are drawn from it and the way in which those inferences are drawn" (GONÇALVES, 2019). By creating a system which is transparent by design, data controllers and processors have more control to identify flaws.

However, Gonçalves (2019, p. 4) stresses that introducing a risk-based approach in data protection regimes raises some concerns. She argues that at "the end of the day, too much will depend on how the data controllers will interpret and fulfil their responsibilities under the GDPR", as the GDPR, and the same applies to the LGPD, delegates to controllers the ultimate reasoning on whether a data processing might be riskful or not. Although both legislations have specific commands regarding the elaboration of data protection impact assessments (DPIAs), which should be undertaken where a given processing is likely to pose high risks for the rights and freedoms of data subjects, the leverage of such a risk is yet a prerogative left to the controller, who is not always the most capable of measuring the impact of its activities.

As we can see, the LGPD and the GDPR do have provisions drawing attention to providing different safeguards according to the level of risk of a data processing activity. At the same time, however, the risk-based approach is only *one* aspect of the GDPR among others (WP29, 2014), and we should argue the same for the LGPD. After all, the rights, risks and legal grounds provided by these laws are applicable for any kind of data processing, and not just to those with a high risk. High risk data processing activities are supposed to take place with stronger procession measures, but this does not mean that lower risk processing is free from compliance.

For this reason, the LGPD's risk provisions are closer to the way that Baldwin and Cave (2021) proposed in the GRID framework that risk levels shall be seen seen as one of the elements composing a regulatory strategy, and not as its only determining factor.

ANPD's Resolution provides for more detail by establishing that its enforcement will be proportional to the risks of the data processing and the behaviour of the regulated agents, which includes the protection measures for compliance (Article 17, IV, Resolution). It also provides for the adoption of a biannual "Map of Priority Themes", which sets ANPD's agenda for studying and planning purposes (Article 21, Resolution) based on, among other actions, the risks involved in a data processing (Article 22, Resolution).

Higher risk processings can also trigger the need for carrying out Data Protection Impact Assessments (DPIA), which will discuss in a deeper detail later on. Considering that the aforementioned Article 20, LGPD, addresses specifically profiling activities that might lead to discrimination, it would not be a surprise if the ANPD would require DPIAs for machine learning systems responsible for high stakes data processings.

#### *5.3.3. Networked governance*

Following Braithwaite's contributions on the inclusion of other actors in the regulatory enforcement beyond regulator and regulated entity, we now turn towards understanding to what degree can we create such a framework under the Brazilian data protection regime.

The LGPD allows for the involvement of non-state actors in enforcement in different ways. One of them is through an obligation that the norms drafted by ANPD should all be preceded by public consultations for the ANPD to listen, at least in theory, to the demands of different interest groups (Article 55-J, §2, LGPD).

It also provided for the creation of the *Conselho Nacional de Proteção de Dados* (National Data Protection Council - CNPD), a multi-stakeholder board formed by representatives of state institutions, civil society, companies, scientific institutions and working unions (Article 58-A, LGPD) with a mandate for providing advice for the Authority (Article 58-B, LGPD).

Nevertheless, it is still quite challenging to put into practice, at least in the Brazilian scenario, Braithwaite's (2006) ideal of giving non-state actors a role of almost a regulator. And that for two main reasons.

First, despite the participatory framework proposed by the LGPD, that allows other actors to influence the adoption of new rules by the ANPD through public consultations and through the CNPD, it is not yet clear to what degree the proposals of these stakeholders will be taken into consideration by the Authority in the drafting of its regulations. Considering the yet recent establishment of the ANPD, which is yet in its second year, Brazil still needs some more time to comprehend how effective such participation will be.

Another point of concern relates to how effective social participation will be in the enforcement of the LGPD, at least at an institutional level, is the fact that it is not always that the Authority includes civil society organisations in consultation. One example was a 2021 consultation about a regulation about data protection impact assessments, where no representative of civil society was allowed by the ANPD to participate (COALIZÃO DIREITOS NA REDE, 2021b).

With regard to the CNPD, the fact that it is presided by a member of the Presidency of the Republic — thus part of the government — and that this member has the power to unilaterally call, suspend and postpone meetings, may end up influencing the effectiveness of the Council's work in themes that go against the government's agenda (Article 3, I, Resolution) (BRASIL, 2022). Moreover, ordinary meetings are to take place only three times a year, which may be seen as a low number when comparing to other data protection boards worldwide, even with different competences from the CNPD (Article 6, Resolution). The European Data Protection Board, for instance, only in the first seven months of 2022 had already had ten plenary meetings (EDPB, 2022).

The Resolution also provides not very elucidative ways for improving other stakeholders' participation. The only element that might be read as a way for promoting networked governance is through the participation of third-party intervenors in administrative proceedings pushed forward by the ANPD (Article 49, Resolution), something which is already a praxis in Brazilian law as a whole. The criteria for third-

party intervenors admission is the relevance, specificity or social repercussion of the theme under analysis in the administrative sanctioning proceeding.

In this sense, although theoretically possible, it is hard to say what are the practical chances that non-state actors, especially civil society, effectively assume a position of co-regulators in Brazil, as Braithwaite (2006) seems to suggest.

In this sense, in order for the ANPD to meet the goals of its agenda in implementing responsive regulation, it is crucial that it include these other actors more effectively in the development of policies related to data protection to obtain their support in the LGPD's enforcement. This is of particular importance considering the low human and financial resources of the Authority, which was initially established without an increase in government expenses (SENADO, 2022b).

Nevertheless, even without specific support from the Authority, civil society has already been active in the oversight of the LGPD. An example happened when the Ministry of Economy announced a partnership with an association of banks, the ABBC, to give it access to a public database with biometric data from millions of Brazilians without detailing to which extent it did not violate the LGPD. Members of the *Coalizão Direitos na Rede* thus rang the alarm to the Authority and to the Federal Public Ministry to take action, and also used their social media accounts to vocalise the issue to society (COALIZÃO DIREITOS NA REDE, 2022).

Other members from the Coalition are at the forefront of a strategic litigation action that was successful in reaching a judicial order determining that the company administering São Paulo's subway system to suspend the use of face recognition in subway stations (IDEC, 2022). NGO's actions also include participations in Supreme Court cases as *amici curiae* (LABORATÓRIO DE POLÍTICAS PÚBLICAS E INTERNET - LAPIN, 2020; COALIZÃO DIREITOS NA REDE, 2021a), in challenging ANPD decisions (COALIZÃO DIREITOS NA REDE, 2021b), among others.

Moreover, in the field of machine learning, organisations from Brazil and other countries in South America gathered to demand more transparency regarding technical aspects about the use of facial recognition systems by government agencies, such as its accuracy rates, and also about administrative issues such as how they are being ac-

quired by the administration (ACCESS NOW, 2021). The action is an example of cross border network, a move beyond national territories to detect data protection violations.

Nevertheless, to optimise the enforcement of the LGPD through networked governance, it is fundamental that the ANPD brings these actors closer in the oversight of data controllers and processors in their compliance. This should happen not only by allowing more space for dialogue and participation, but also by supporting civil society organisations and academic institutions in receiving fund, including from the government. This is a key action especially when taking into consideration the potential of external researchers in supporting regulators, as happened in the COMPAS case we mentioned above (LARSON, 2016).

Beyond these, other communication channels should be opened by the regulator not only with organisations, but also with victims of data processing, including by machine learning systems. One way for putting this in practice would be, for instance, allowing for citizens who may be affected by face recognition systems, especially considering the racial biases we pointed earlier, to provide their views on the risks of these tools, and thus influence decision-making on whether they should be or not adopted by their local governments.

A final remark should be made regarding the figure of the *encarregado*, which finds a parallel in the GDPR with the data protection officer (DPO) who is the "person appointed by the controller to act as a channel of communication between the controller and the data subjects and the supervisory authority" (Article 5, VIII, LGPD). The *encarregado* has the role of creating a trusted environment for enhancing cooperation under a responsive regulation rationale by monitoring compliance of his/her organisation to the LGPD and by issuing recommendations, while being the point of contact of the data controller or processor with other stakeholders (IRAMINA, 2020, p. 107). In a networked governance, thus, *encarregados* can play a crucial role in representing the node of the regulated entity during dialogues with other nodes of the enforcement network.

## 5.4. Machine learning and responsive regulation

Having presented the main fundamentals of the responsive regulation theory and how positive regulation contributes to it, as well as how it can be applied under the scope of the LGPD, the question we now aim to respond is how to use the tools they provide to push forward for more transparent machine learning systems that process personal data.

The reduction of information asymmetries is a fundamental aim of regulation. It is a necessary condition for authorities to obtain more bargaining power to design the suitable strategy to regulate an agent. Without knowing enough about how an organisation performs its functions, a regulator is hardly able to understand its business to effectively assess compliance levels or determine whether the entity is to be considered virtuous or not. This includes, depending on the actor, analysing whether it has put in place sufficient safety measures to protect consumers and employees, whether it is following standards for reducing carbon emissions, among others (AYRES, BRAITHWAITE, 1992; BALDWIN, CAVE, LODGE, 2012).

It is for exactly this reason that a crucial part of regulatory enforcement for the positive theory, which we assessed earlier, is obtaining information on non-compliant and undesirable behaviour. In the field of data protection, ML transparency is a crucial feature for reducing information asymmetries involved in personal data processing supported by automated decision means.

In the previous chapters, we saw that the transparency of ML systems is context-dependent. As such, the type, amount and language of the information provided about the system and its environment depend on issues like the target-group of the information, who is deploying the systems, in what contexts and so on. Making regulation flexible is thus of utmost importance.

### *5.4.1. ML transparency: the case for flexibility*

Responsive regulation is regarded as a reaction against views that advocated for regulatory strategies based strictly in the dichotomy command and control/self-regulation, in which only one of these strategies should be applied by the regulator.

We have already argued that the pure self-regulatory strategies, even when based on ethic principles, are insufficient to tackle the impacts of many of ML systems, including when processing personal data. A command and control strategy, on the other hand, would also be problematic to effectively regulate these systems if applied by regulators as the only strategy. But some thought over it is suitable at this point.

Command and control is based on the notion that law can be drafted in a way that imposes fixed standards with immediacy and at the same time prohibits and penalises any activity that does not conform to them (BALDWIN, CAVE, LODGE, 2012). It is thus based on the notions that (i) the threat of sanctions are *per se* enough to promote compliance and that (ii) the legal system expresses itself through coercion (ARANHA, 2019).

Bringing it to our context, a command and control strategy could demand from the legislator or regulator to establish the most exhaustive possible rules for addressing the transparency of ML systems, and, if these commands are not followed, the Authority would be in a position to punish the offender. However, we saw that ML transparency is context dependant, and its effectivity is affected by a myriad of factors that would hardly work if strictly determined.

For credit bureaus, e.g., probably counterfactual explanations would be an interesting tool to deliver justice, as they would allow for an individual to understand what factors about her life are being more crucial for the denial of credit by disclosing how changing specific inputs would allow her to reach a different output.

Periodic data protection impact assessments (DPIAs) and reports could also be suitable to show what are the profiles that are having more credit requests refused or how diverse are the staffs taking the decisions related to the ML systems they operate, and thus help assess potential discriminatory outputs.

Differently, image recognition systems could be evaluated through assessing training databases in order to find potential racial or gender-based discriminatory biases. Issuing reports on accuracy rates or on what were the criteria used for categorising images in its training dataset, for instance, could also be useful.

Moreover, in the COMPAS case we mentioned in Chapter 4, an algorithmic system used to assess a criminal defendant's likelihood of becoming a recidivist, what led to a conclusion by external researchers that it had racial biases was the access to the database with the scores it had assigned to individuals, and not to the code itself (LARSON et al., 2016). COMPAS and other recidivism systems are another case of system that is almost incapable of avoiding automatising racist practices, taking into consideration the fact that they are being used in an environment which is already one of the main representations of structural racism, which is the criminal system itself (ALMEIDA, 2019).

The examples above show how different information might be necessary to allow for detecting flaws in ML systems. In this sense, making use of command and control strategies, that aim at clearly and exhaustively defining parameters, would probably be insufficient to provide for qualified transparency. Even if the different techniques for transparency provision exemplified above were translated in legal rules, it would be hard to define in detail for which systems would each of them apply.

To say the least, this is due first because the systems might demand a series of experiments to achieve a reasonable level of understandability. After all, if scholarship has reached very few consensus on what are the best ways for explaining ML applications, it is hard to say that regulator and data controllers will reach an agreement so easily.

Second, because new ML applications, as well as transparency-promoting methods, are being developed at such a high speed that it would be hard to imagine the regulator managing to create rules capable of matching them in due time. Regulators would thus always be one (or many) steps behind innovation, and the rules they provide will be destined to be always outdated.

### *5.4.2. Regulatory pathways*

For this reason, flexibility in defining strategies based on responsive regulation might be more effective in pursuing the adaptability to regulate machine learning transparency (AYRES, BRAITHWAITE, 1992). ANPD could, for instance, demand from a credit bureau to develop a transparency plan describing how it will issue periodic reports about its systems' activities, how it aims to address eventual biases that already exist and how it will create a platform for consumers to access counterfactual explanations. It can ask the same for a hospital that uses biometric data to authenticate its patients, while also assessing data quality by auditing the databases the institution uses to train its systems.

If these measures are seen as enough by the Authority, it will validate them and make it applicable to the credit bureau from then on. If the credit bureau or the bank do not follow them, the Authority will be in a position to punish it for not following its own rules that were negotiated with the regulator. If these measures are proven insufficient or misleading, the Authority could either order the controllers to amend them, amend them on its own and even, if the agent is perceived as drafting the rules with bad faith in order to hide information, punish it for that and make an audit on its own.

Such rules can be designed, within the Resolution, under the scope of a conformity plan, of which not acting in accordance with it will be an aggravating factor for potential sanctions (Article 26, Resolution).

This strategy would be closer to what Ayres and Braithwaite (1992, pp. 105-6) call *enforced self-regulation*.

At the same time that the authors argue that self-regulation can be a very cost effective strategy to adopt, and that insiders can frequently have more capacity to trap wrongdoers in an organisation, they are "not necessarily more willing to regulate effectively. This is the fundamental weakness of voluntary self-regulation".

In these cases, self-regulation can be imposed by the regulator, in a way that the entity is compelled to write a set of rules tailored to the unique set of contingencies it faces, to be subject by the regulator's decision to either approve these rules or send them

back for revision. In this process, external actors such as civil society organisations, affected groups, researchers and others would be encouraged to comment on the proposed rules. If data controllers do not follow the rules established in their rulebooks or fail to provide for regulatory tools to promote transparency that is effective to tackle ML's pitfalls in a way that encompasses the system and the social systems in which they dwell, the regulator should be ready to punish, even with the use of the heaviest sanctions provided by the LGPD (AYRES, BRAITHWAITE, 1992).

Of course, and this is part of the general logic of the responsive regulation theory, this strategy will not be applicable to any circumstance, and has its own flaws. On the one hand, it is an interesting approach since its rules would be tailored to match the company's particularities and would adjust more quickly to changing business environments. On the other hand, regulatory agencies would have to bear costs of approving a vastly increased number of rules each year. It is up to the regulator to thus understand to what degree it is a good strategy, preferably opening for other stakeholders's opinions, including from those who might be affected by these systems and who are in a vulnerable position to make their rights enforceable.

Moreover, there should always be flexibility for the ANPD to impose LGPD's heaviest sanctions or to adopt strategies more focused on command and control in cases where the regulated entity has a history of non-compliance or even disrespect for institutions. As we saw in Baldwin and Cave (2021), the combination of high risks, high capacity and ill intentions can be a perfect storm for the tougher forms of enforcement.

Mechanisms for reversal of the burden of proof can also be put into use by a regulator if it identifies law violations. According to Article 50 of the Resolution, the "offender is responsible for proving the facts he alleges, without prejudice to the duty attributed to the competent organ for investigation". This could be similar to reversing the burden of proof when information asymmetries are not sufficiently addressed with the information that reaches the Authority, or when such information is so overwhelming that ends up being impossible to understand.

All of this will depend, again, on the context. Robert Baldwin (2021), in an online lecture for students from the University of Brasília, affirmed that in cases that when opacity does not allow one to scrutinise the underpinnings of what the regulated entity argues, and/or when the agent was not cooperative with the authority, reversal of the burden of proof would be possibly a good technique to assess compliance. In this case, the regulator would be in a position to ask "convince me that you are behaving properly".

### *5.4.3. Data Protection Impact Assessments*

Beyond these potential instruments, another which might be included within the scope of one of a strategy for promoting transparency about ML systems can be Data Protection Impact Assessments (*Relatório de Impacto à Proteção de Dados Pessoais* - DPIA). They are provided by both the LGPD and the GDPR for accessing information about a personal data processing, and are supposed to be developed by data controllers in different situations.

The LGPD describes the DPIA as the "documentation from the controller that contains the description concerning the proceedings of the personal data processing that could pose risks to civil liberties and fundamental rights, as well as measures, safeguards and mechanisms to mitigate said risk" (Article 5, XVII, LGPD). They should contain at least a description of the types of data collected, the methodology used for collection and for ensuring the security of the information, and the analysis of the controller regarding the adopted measures, safeguards and mechanisms of risk mitigation (Article 38, *par. un,*, LGPD).

The Brazilian legislation does not provide for a list of cases in which the DPIA is mandatory. However, it establishes some examples of situations in which the ANPD might request its carrying out to the data controller. These are:

1. Personal data processing activities for purposes of national security and law enforcement (Article 4, §3, LGPD);

2. Personal data processing activities based on legitimate purposes (Article 10, §3, LGPD);

3. To data controllers in the public sector (Article 32, LGPD);

4. Personal data processing activities involving sensitive data (Article 38, LGPD).

Kaminsky and Malgieri (2019), when assessing DPIA rules in the GDPR, affirm that these assessments can be an important layer of transparency for ML systems. Although the GDPR's provisions on DPIAs hold many particularities that are not included in the LGPDs, some ideas of the authors can be very well suited for the scope of the Brazilian law.

Interpreting how the model of governance proposed by the GDPR would reflect on ML transparency, Kaminsky and Malgieri (2019, p. 5) first argue that the GDPR proposes a system of a multi-layered explanation rationale, whereby "[i]ndividuals have a right to both a system-wide but detailed description of the logic of an algorithm (Arts. 13, 14, 15), and more specific insights on individual decisions taken". In this sense, the more intrusive or riskier a system is, further information it would have to disclose so as to allow for individuals to effectively exercise their rights.

According to the authors, DPIAs and, more specifically, Algorithmic Impact Assessments (AIA), should play a crucial role in allowing for transparency enhancement. Similar to the DPIA proposed by both LGPD and the GDPR, AIAs would function as a tool to achieve algorithmic accountability by assessing artificial intelligence (including ML) systems' impact on individuals' and groups' rights (KAMINSKY, MALGIERI, 2019, p. 13). As AIAs are not specifically prescribed under the GDPR (nor the LGPD), the authors present their idea inspired by DPIAs, and, due to some diverse rationales inherent to each of these two instruments, further regulation would probably have to be yet designed in both Europe and Brazil to allow for AIAs.

DPIAs, when applied to assess ML automated decision making systems, would work to analyse the degree of risk that a data processing poses to natural persons, and provide for the necessary actions for reducing or avoiding these risks. They would thus be a first disclosure of information regarding how the system being assessed works, in

a form of what they call “monitored self-regulation”. Consequently, these assessments should trigger data controllers to come up with concrete ways to mitigate the risks it might pose (KAMINSKY, MALGIERI, 2019, p. 16).

For this reason, they can play a crucial role in a responsive regulation dynamic. In case the ANPD, for instance, calls for a data controller to disclose information on a specific system that has presented data protection violations, the DPIA might be used as an evidence that the controller has adopted every measure at his or her disposal to mitigate risks. It might, hence, help avoid a massive escalation of sanctions in an enforcement pyramid or gradient due to the controller’s good faith.

Although neither the LGPD nor the GDPR provide for an obligation of disclosure of DPIAs, their publication is recommended by the authors. In this sense, they argue that, in case controllers disclose at least a summary of DPIAs and AIAs, it can include a first layer of explanation regarding these systems that informs external stakeholders about the general logic of the system. Such a layer could be further complemented by explanations on a group-level, for analysing how an algorithm might impact particular classes of individuals, or particular locations, and, further, on an individual-level (KAMINSKY, MALGIERI, 2019, p. 27).

As a result, DPIAs and AIAs could work as one effective tool among others for allowing regulators to understand how ML systems deployed by data controllers affect rights and liberties of individuals and groups and the degree of risk that they pose. Such understanding is paramount for, first of all, assessing compliance of the controller to the LGPD and the Brazilian legal system as a whole through analysing whether the regulated entity adopted reasonable mitigating actions has acted or not in good faith. Secondly, to identify whether more information is necessary to allow for the comprehension of the ML system and, hence, enable an individual or group to exercise the rights provided for in the LGPD.

#### *5.4.4. Other techniques*

In the context of ML transparency, beyond the strategies and tools already mentioned, we can think of the inclusion of information regarding these systems in datasets managed by authorities and the drafting of technical documentation and impact assessments with details about the functioning of the system, tools provided, for instance, under the European Union's (2020) AI Act proposal. It is worth noting, however, that the EU's proposal has been criticised for many reasons, including for the fact that its original draft fails to empower affected individuals (EDRi, 2022) and has insufficient transparency provisions (ALÍ, YU, 2021).

Another valid tool is the emission of notices informing persons that they are interacting with ML systems or with ML-generated content. An example can be found under Article 14 of the Chinese legislation about Provisions on the Administration of Deep Synthesis Internet Information Services (CHINA, 2021), that aims to regulate deep fakes. The provision determines that "[w]here deep synthesis service providers provide […] deep synthesis services, they shall identify the deep synthesis information content in a conspicuous way to effectively alert the public about the synthetic nature of the information content", which applies to the ones interacting with the content.

Its Article 12, on its turn, provides that deep fake creators and users shall "inform and obtain the independent consent of the entity whose personal information is being edited".[16] This latter obligation is more easily framed within the scope of LGPD and ANPD's enforcement, as it directly involves the processing of personal data of the subject whose deep fake is about. It is important to note, however, that in many cases the creation of deep fakes is made within the scope of artistic works, which, as we mentioned previously, fall out of the scope of the LGPD.

It is fundamental to note that, in each of the myriad of strategies that we highlighted, and also of the many other strategies that also possible to adopt under the responsive regulation model, it is fundamental for the regulator to ask the questions we posed in Chapter 3. They are a key roadmap for the ANPD and other stakeholders to assess to what degree the information they have is sufficient for achieving its goals or, if they

16 The translations for this legislation were extracted from CHINA LAW TRANSLATE. Provisions on the Management of Algorithmic Recommendations in Internet Information Services. 2021. Available at https://www.chinalawtranslate.com/en/deep-synthesis-draft/. Last accessed 28 July 2022.

need further data, what exactly is the information they need, to whom, for what and to which stakeholder.

In all of the arrangements highlighted, the inclusion of other stakeholders is an important feature to improve enforcement. As we saw in Braithwaite (2006), regulators lack the necessary resources to take care of a whole market. When talking about data protection, this is particularly challenging since the processing of personal data is now part of possibly every industry in the world, and the tools applicable are the most varied.

Including other actors in this context can support enforcement (ALÍ, YU, 2021). Civil society organisations, researchers, affected communities, standard bodies, can all be included by the regulator not only in the design of norms, as in strategies involving enforced self-regulation, but also in the monitoring of the activities of data controllers. Nevertheless, this should come side by side with policies that enhance institutional and financial support to these groups in order to gain the expertise and workforce necessary to support effectively in the enforcement of the LGPD.

Finally, providing for a safe and encouraging environment for whistle-blowers to leak cases of violations of the law is also an element that might increase revelations about the deployment of ML systems and how they are affecting society and the natural environment. A great part of the largest scandals of the last years involving data protection abuses and ML systems' impacts came from leaks brought to light by insiders, such as the Snowden revelations, the Cambridge Analytica scandal and the Facebook Files, to say the least. Creating room for further protection and opportunities for the future whistle-blowers is thus fundamental for us as a society to better understand the effects of the technologies we are dealing with.

#### *5.4.5. Further thoughts*

Many other strategies would be possible under the model of the responsive regulation theory, and can either include the use of pyramids and gradients, or ignore them by escalating straight to heavier sanctions or strategies. Their design will all depend on

issues such as the context, the risk of the application, the type of information necessary or the willingness of the agent to cooperate.

However, it is necessary to have in mind, as we mentioned a few times in the pages above, that transparency is one of the means towards a myriad of possible ends, and which sometimes might not even be a necessary object towards which it is worth spending regulatory resources.

Providing effective means for the exercise of other principles and rights in the LGPD including free access, right to erasure — including through the erasure of data in ML models or the erasure of ML models themselves — and data portability is thus a fundamental piece of the puzzle in order to achieve the effective enforcement of data protection rights in a way that go beyond transparency (VEALE, EDWARDS, 2017).

The effective compliance with the LGPD can also be enforced through the strengthening of privacy by design tools that aim to embed the best technical and governance practices personal data processing so as to conduct it in the most privacy and data protection enhancing manner, thus protecting data from any form of illicit or inadequate processing (Article 46, LGPD). This includes the carrying out of Data Protection Impact Assessments, which, by identifying potential risks, allows a data controller or processor to adopt better tools to protect the information or even give up of the data processing at all (VEALE, EDWARDS, 2017).

For these reasons, an assessment of the context in which the processing of personal data with the support of a machine learning system is necessary in order to understand, first, whether more transparency is indeed indispensable or whether a solution can be reached without spending resources in this direction; and, if further information is imperative, how the access and the information itself might be more effective in delivering understanding to its recipient. At this point, going back to the set of questions we came up with is fundamental to deliver the most qualified transparency possible.

## 6. CONCLUSION

*Through the veils*

This journey, this *travessia*, started with an unrest. How come humanity is adopting massively, in ever increasing rapidness, a set of technologies that are known to pose major social and environmental impacts that we are not allowed to scrutinise because that would reduce profits and interfere in consolidated power dynamics?

This unrest grew as reports on machine learning-led automatisation of discriminatory biases present in society for centuries proliferated under a veil of opacity and cynicism typical of a system that reproduces itself through the maintenance of unjust structures of power. These situations are perceived in how face recognition systems deployed by the police in the cities of Salvador and Rio de Janeiro target mostly black and poor individuals who are already historically the ones persecuted and exterminated by the state. They are also perceived in Italy, where cities like Rome and Bologna are testing ML systems to score citizens based on their social behaviour (REMIX, 2022), and in the plans of different European governments, who aim to deploy ML (and other AI techniques) in their borders to control those whom they judge as unworthy of making part of their wonderland (STATEWATCH, 2022).

With the opacity of machine learning systems being so common in both the public and private sectors, we investigated in the previous pages issues such as what are these technologies about, for what are they being deployed and what is their relationship to data. Other aspects included how opacity in these systems expresses itself, whether it is indeed problematic or not and, if so, how to tackle it.

In this sense, we assessed the arguments that have been used to impose resistance against more machine learning transparency, or, better still, qualified transparency. Arguments frequently used to avoid this, so to say, opening of the black box, are that releasing information on algorithms may lead bad-faith actors to game the systems and that such information is legally protected as a trade secret (BAYAMLIOGLU, 2018). Many have also alleged that it is an impossible task to completely understand

ML applications due to their complexity, which make them opaque even for their own developers, as described by Veale and Edwards (2017).

Against this background, ML continues to be used to support many contestable predictions and decisions frequently based on technosolutionist mindsets that take technological inevitability — which we regarded as a fallacy (ZUBOFF, 2019) — as an imperative. By doing so, the ones deploying these systems end up some times discriminating against vulnerable individuals pertaining to lower-income groups, migrants, ethnic minorities, indigenous, afro-descendants, LGBTQIA+, as well and many others who have historically been marginalised from society.

Rendering these systems understandable can thus be a key instrument to allow not only for oversight and accountability but also to help society gain knowledge on how machine learning is helping shape society. This process of incrementing understanding should not only relate to making their mechanical functionings understandable, but also the contexts in which they are designed, built, applied and discarded.

To adopt this holistic approach to transparency is fundamental to grasp how these systems are not only reproducing already existent social discriminatory biases but also automating them. Moreover, they will help us understand what are the labour and environmental impacts that ML systems are posing to the world in the form of phenomenons such as ghost work, ML-controlled work, energy consumption, mining and etc.

*Transparency's limitations*

However, transparency is not something easy to be provided, and is far from allowing, per se, for an effective form of control. Seeing, having access to a given object, is not by itself enough for granting the necessary knowledge to render it accountable, and this applies not only for algorithms, but for many other social systems. Despite the fact that the access to a system's inner workings can indeed provide insight and spur further investigation, "significance and power is most revealed by understanding both its viewable, external connections to its environments and its internal, self-regulating workings" (ANNANY, CRAWFORD, 2016, p. 978).

Transparency has thus many limitations, which include issues related to lack of trust on the information provided in certain contexts, a tendency of transparency privileging seeing over understanding, and strategies for occluding through transparency, by means of informational overload. For this reason, not necessarily understanding a system allows one to effectively change the way it works or even make it not be used at all.

ML technologies should thus be seen as "sociotechnical systems that do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans" (ANNANY; CRAWFORD, 2016, p. 974). Transparency is thus "only the beginning of this process" (EDWARDS; VEALE, 2017, p. 41), or one out of many other means for achieving effective accountability, specially concerning ML systems (DOSHI-VELEZ & KORTZ, 2017). Further, it can be seen as valid only to the extent to which it allows for effective action and reflexivity about how ML helps perpetuate exclusion (D'IGNAZIO, KLEIN, 2020; SILVA, 2022).

Framing the best ways for delivering transparency regarding ML systems finds limitations not only on a conceptual level, but also on a technical one. The operations of a system involve a series of layers such as code, data, inputs, outputs, that provide for a range of different information that can overload an individual trying to see through them. In this sense, one should always have clear in mind what it is that can be useful to understand in the system in order to achieve a specific goal.

As such, the level of understandability of a system depends on issues that include how the information is made available, the target group of such information, what are the goals involved in the disclosure, the type of system being assessed and the risks posed by it are necessary questions that one needs to make in order to frame a strategy for demanding information about a system. Transparency about ML systems is thus highly context-dependent, and any regulatory approach to promote it should be flexible enough to cover the particularities of different systems, developers, deployers, end-users and affected communities.

*ML transparency and data protection*

We thus started a journey to understand how these issues were covered by data protection laws in Brazil and the European Union, respectively the LGPD and the GDPR. The similarities of both regimes reside, among other aspects, in the scope of application of the acts' provisions, which includes any activities involving the processing of personal data (except for some exceptions which we will see below), in the existence of lawful grounds for processing data, as well as of general principles and rights.

Those choice for looking at the issue of ML transparency through the lens of data protection arose after realising that many of the most problematic ML systems that were mentioned throughout this work process personal data. In this sense, without having laws specifically focused on regulating these technologies approved in these jurisdictions, data protection regimes could provide us with guidance to tackle the transparency questions we highlighted.

One can read the LGPD and the GDPR as designed to create a barrier against massive personal data processing by both public and private entities, irrespective of their sizes. As the access to data sets and new technologies capable of processing extensive and sensitive amounts of data and of extracting patterns from them (including machine learning) was democratised, now many small companies can also take part in the surveillance capitalism that we talked earlier in the Introduction (ZUBOFF, 2019).

Within these laws, we identified that a large discussion has taken place in the last years over whether there is or not a right to explainability in their scope, mostly focusing on specific provisions related to automated decision-making regulation. However, after assessing the multiple sides of this debate, this work considered, inspired by Kaminski (2018), that the LGPD and GDPR both provide for rights and obligations regarding the provision of information that go beyond the discussion restricted to the so-called right to explanation. These include provisions establishing transparency principles, rights of access, rules for auditing and so on. Transparency should thus be read in a holistic way, considering that it is a crucial part of these laws not only with regard to personal data processing activities with the aid of ML system but any personal data processing in general.

*The role of responsive regulation*

Considering these many legal, technical and social peculiarities involved in rendering ML systems transparent, this work tackled how, from a regulatory perspective, should the enforcement of these provisions take place.

Based on the findings from the previous pages, a regulation of ML systems, especially for the enforcement of the framework of the LGPD with regard to transparency, should take into consideration two main features. First, it should be capable of allowing for sufficient malleability of the regulator in order to implement policies that are suitable for addressing different ML systems governed by different agents in different contexts. Second, it should allow for a constant dialogue not just between regulator and regulated entities, but also other interested stakeholders, including civil society organisations, researchers, auditors, representatives of affected groups, and many others.

With that in mind, this work assessed the suitability of the theory of responsive regulation to guide the enforcement of the LGPD with regard to ML systems. This choice was motivated by both the idea of flexibility intrinsic to the theory, but also because the National Data Protection Authority - ANPD has already signalled that its enforcement would be guided by this theory (URUPÁ, 2021; BRASIL, 2021b).

The responsive regulation theory is aimed at overcoming the simplistic regulate-deregulate debate by claiming that, for a regulation to be effective, it should create rules that incentivise a regulated entity to voluntarily follow them, through strategies that range from granting awards and quality seals to imposing harsh punishments (AYRES & BRAITHWAITE, 1992). All this should be made in an environment of constant dialogue between regulator and regulated, in a way that reduces information asymmetries between these actors and help align the interests of regulated actors and society (ARANHA, 2019) by shaping the behaviours of actors in the regulatory game (AYRES & BRAITHWAITE, 1992, p. 44).

Responsive regulation is, hence, mainly about finding the right balance between punishment and persuasion in order to make regulation effective. Its theorists argue that, when regulators adopt strategies based strictly on punishment, their actions undermine

the good will of actors when they are motivated by a sense of responsibility. On the other hand, however, when the strategy is based exclusively on persuasion and self-regulation, state action will probably be exploited when actors are motivated exclusively by economic rationality (BRAITHWAITE, 1985).

This leads us to the notion that the kind of regulatory strategies to be adopted by regulators that aim to follow the responsive regulation theory is — similarly to ML transparency — context-dependent. This means that the characteristics of the regulated agent, its economic and political power, the peculiarities of the technology, the market in which it is being applied, the consumers using it, the affected individuals, among others, have all the ability to influence the strategies adopted by the regulator.

In understanding persuasion and punishment as interdependent and complementary in the design of regulatory strategies, the responsive regulation theory combines regulatory incentives and intrusive measures as means for the concretisation of regulatory objectives under a format of enforcement pyramids. They allow for regulators to escalate towards more intrusive constraints as offenders break more rules or act viciously and also to de-escalate when regulated entities improve their compliance culture through time (AYRES, BRAITHWAITE, 1992, p. 35; ARANHA, 2019, p. 113). These pyramids, as we saw, may be complemented by the Positive Regulation Theory's GRID, an enforcement framework in form of a gradient that considers the regulated entity's capacity, intention and risks as elements for influencing oversight and the imposition of penalties by the regulator (BALDWIN, CAVE, 2021).

Moreover, as the responsive regulation theory was created in Australia, we assessed the arguments brought by John Braithwaite (2006) on how to apply responsive regulation in the Global South, which holds particular characteristics. He draws attention to what he refers as a smaller regulatory capacity in the Global South and to a supposedly minor oversight by NGOs and social movements to mobilise as potential inhibitors of regulatory success (BRAITHWAITE, 2006, p. 885).

To tackle these issues, the author proposes that enforcement is put into practice with the cooperation of other actors in the regulatory framework in order to be more effective. That includes NGOs, audit offices, ombudsmen, courts, public service commis-

sions, self-regulatory organisations, among others. They should be called by the regulator to be responsible for the oversight of legal rules, holding not only regulated entities accountable, but also everyone else within the regulatory dynamics, be it a state or non-state actor. (BRAITHWAITE, 2006) This is what Braithwaite (2006, p. 886) would call "nodes of networked governance".

There are many possible ways that the data protection framework in Brazil may allow to apply the responsive regulation theory when enforcing transparency obligations related to the use of ML systems. These include scalable sanctions, flexibility in designing strategies for different regulated entities, receiving contributions from multiple stakeholders, development of guidelines and educational actions, preventive measures, development of conformity plans, among many others. However, further political and regulatory will is necessary to put the possibilities provided by the LGPD into practice, especially when considering the possibilities of putting a networked governance into practice.

*A post-scriptum*

The regulatory paths that we discussed in this work aim to open a glade at the core of such an unknown field as machine learning regulation. It is easy for us to get lost in the discussion of such a complex and polyvalent set of technologies. However, perhaps the hugest challenge we might be facing as a society today, especially in Brazil and other countries in the Global South, might be to understand to what degree are these technologies serving our communities as tools for social cohesion, and not to perpetuate values that further strengthen power dynamics.

Racism, colonialism, misogyny, are just some of the issues that are being perpetuated, automated and multiplied through the use of these tools. We need thus to consider why is it that we are adopting these technologies, for what goals, to solve what issues, based on what beliefs. And in this process, reclaim our prerogative to understand them if they are affecting us.

We need to take a step back to reflect on the regulation of these technologies in terms of creating and deploying applications that *transform* power structures, and not just avoid worsening them. While so many are talking about de-biasing systems, we need to start conceiving technological advancement as only desirable to the extent that it is used to *promote* equality, to balance power, to give space and opportunities to those who have always been excluded from them.

Transparency is one of the potential paths towards this goal, but is far from being the only one. It is crucial to determine what are the tools that we as a society want and what are the ones we should avoid. We, within our communities, are the ones who should have the power to decide what is the technological future that we want.

**BIBLIOGRAPHY**

ACCESS NOW. **Made Abroad, Deployed at Home**: Surveillance Tech in Latin America. 21 August 2021. Available at https://www.accessnow.org/cms/assets/uploads/2021/08/Surveillance-Tech-Latam-Report.pdf. Last accessed: 26 July 2022.

ADAMS, Chris. How much CO2 can you save when you remove ad-tracking from news sites? **Chris Adams Blog**, 27 May 2018. Available at https://blog.chrisadams.me.uk/posts-output/2018-05-27-how-much-co2-can-you-save-when-you-remove-ad-tracking-from-news-sites/. Last accessed 14 July 2022.

ALLEN, Robin; MASTERS, Dee. Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. **ERA Forum**, [*s. l*], v. 20, n. 4, p. 585-598, Mar. 2020.

ALÌ, Gabriele Spina; YU, Ronald. Artificial Intelligence between Transparency and Secrecy: from the EC white paper to the AYA and beyond. **European Journal Of Law And Technology**, [*s. l*], v. 12, n. 3, p. 1-25, 28 dec. 2021.

ALMEIDA, Silvio. **Racismo estrutural.** São Paulo: Pólen, 2019.

ALTER, Charlotte. How Fixing Facebook's Algorithm Could Help Teens—and Democracy. **Time**. 05 October 2021. Available at https://time.com/6104157/facebook-testimony-teens-algorithm/. Accessed on 18 November 2021.

ALVAREZ; MARSAL et al. LGPD no mercado brasileiro. S.L: Alvarez & Marsal, 2021. Available at: https://www.alvarezandmarsal.com/sites/default/files/2021-11/E-book%20LGPD%20no%20Mercado%20Brasileiro.pdf. Last accessed: 6 jun. 2022.

ALVES, Fabrício da Mota; VALADÃO, Rodrigo Borges. ANPD: Agência reguladora ou autoridade reguladora independente? **Migalhas,** 7 July 2022. Available at https://www.migalhas.com.br/coluna/dados-publicos/369257/anpd-agencia-reguladora-ou-autoridade-reguladora-independente. Last accessed: 12 July 2022.

ANANNY, Mike; CRAWFORD, Kate. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. **new media & society**, v. 20(3), pp. 973–989, 2016.

ANTUNES LIMA DA FONSECA CARVALHO, J. P. **The Legal Status of the National Data Protection Authority in light of the Regulatory State Theory**: Is there any room for the adoption of the material concept of administrative decentralization in Brazil?. Law, State and Telecommunications Review, *[S. l.]*, v. 12, n. 2, p. 118–132, 2020. DOI: 10.26512/lstr.v12i2.34714. Disponível em: https://periodicos.unb.br/index.php/RDET/article/view/34714. Acesso em: 4 apr. 2021.

ARANHA, M. I. Manual de Direito Regulatório: Fundamentos do Direito Regulatório, 5a ed. rev. ampl, London: Laccademia Publishing, 2019.

ARAUJO, Gabriel et al. Brazil health ministry website hit by hackers, vaccination data targeted. **Reuters**. 11 Dec. 2021. Available at: https://www.reuters.com/technology/brazils-health-ministry-website-hit-by-hacker-attack-systems-down-2021-12-10/. Last accessed: 6 June 2022.

ARTICLE 29 DATA PROTECTION WORKING PARTY (WP29). **Opinion 4/2007 on the concept of personal data**. Brussels: European Commission, 2007.

_____. **Opinion 1/2010 on the concepts of "controller" and "processor"**. Brussels: European Commission, 2010.

_____. **Statement on the role of a risk-based approach in data protection legal frameworks.** Brussels: European Commission, 2014.

_____. **Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679**. Brussels: European Commission, 2017.

_____. **Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679**. Brussels: European Commission, 2018.

ARYA, Vijay *et al.* One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. 2019.

ASGHARI, Hadi *et al*. **What to explain when explaining is difficult?**: an interdisciplinary primer on XAI and meaningful information in automated decision-making. Berlin: Alexander von Humboldt Institute For Internet And Society, 2021.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). **Resolução CD/ANPD Nº 1, de 28 de outubro de 2021**. Aprova o Regulamento do Processo de Fiscalização e do Processo Administrativo Sancionador no âmbito da Autoridade Nacional de Proteção de Dados. Brasília: 28 Oct. 2021. Available at https://www.in.gov.-br/en/web/dou/-/resolucao-cd/anpd-n-1-de-28-de-outubro-de-2021-355817513. Last accessed 27 July 2022.

AYRES, Ian; BRAITHWAITE, John. **Responsive Regulation:** Transcending the Deregulation Debate. Oxford University Press, USA, 1992.

BALDWIN, Robert; CAVE, Martin. **Taming the Corporation**: How to Regulate for Success. Oxford University Press, 1st Ed., 2021.

BALDWIN, Robert; CAVE, Martin; LODGE, Martin. **Understanding Regulation**: Theory, Strategy, and Practice. Oxford: Oxford University Press, 2nd Edition, 2012.

BARRETT, Lisa Feldman *et al*. Emotional Expressions Reconsidered: challenges to inferring emotion from human facial movements. **Psychological Science In The Public Interest**, [*s. l*], v. 20, n. 1, p. 1-68, 2019.

BAYAMLIOGLU, Emre. Contesting Automated Decisions: A View of Transparency Implications. **European Data Protection Law Review**, v. 4, n. 4. pp. 433-446, 2018.

BLACK, Julia. Constructing and contesting legitimacy and accountability in polycentric regulatory regimes. **Regulation & Governance**, [S.L.], v. 2, n. 2, p. 137-164, jun. 2008. Wiley. http://dx.doi.org/10.1111/j.1748-5991.2008.00034.x.

BLACK, Julia; MURRAY, Andrew. Regulating AI and Machine Learning: Setting the Regulatory Agenda. **European Journal of Law and Technology**, Vol 10, Issue 3, 2019.

BORGESIUS, Frederik J. Zuiderveen. Singling out people without knowing their names – Behavioural targeting, pseudonymous data, and the new Data Protection Regulation. **Computer Law & Security Review**, [S.L.], v. 32, n. 2, p. 256-271, abr. 2016.

BOUCHER, Philippe. Une division de l'informatique est créée à la chancellerie "Safari " ou la chasse aux Français. 1974. Available at: https://www.lemonde.fr/archives/article/1974/03/21/une-division-de-l-informatique-est-creee-a-la-chancellerie-safari-ou-la-chasse-aux-francais_3086610_1819218.html. Last accessed 20 Feb. 2021.

BOVENS, Mark. **Analysing and Assessing Accountability**: A Conceptual Framework. European Law Journal, Vol. 13, No. 4, July 2007, pp. 447–468. doi:10.1111/j.1468-0386.2007.00378.x.

BOWKER, Geoffrey C.; STAR, Susan Leigh. **Sorting Things Out**: classification and its consequences. Cambridge, Massachusetts: The MIT Press, 1999.

BOZDAG, E. Bias in algorithmic filtering and personalization. **Ethics and Information Technology**, 2013, 15(3), 209–227.

BRAITHWAITE, John. **Corporate Crime in the Pharmaceutical Industry**. Londres: Routledge & Kegan Paul, 1984.

_____. **To Punish or Persuade:** Enforcement of Coal Mine Safety. Albany: State University of New York Press, 1985.

_____. Responsive Regulation and Developing Economies. **World Development**, v. 34, n. 5, p. 884 – 898, 2006.

_____. The Essence of Responsive Regulation. **UBC Law Review** 44(3), 2011, pp. 475-520. Available at http://johnbraithwaite.com/wp-content/uploads/2016/03/essence_responsive_regulation.pdf. Last accessed 26 July 2022.

_____. Restorative justice and responsive regulation: The question of evidence. **RegNet Research Paper**, No. 51 (revised), School of Regulation and Global Governance (RegNet), 2016.

BRASIL. Resolução CNPD nº 1, de 6 de maio de 2022. Estabelece o Regimento Interno do Conselho Nacional de Proteção de Dados Pessoais e da Privacidade. Brasília, 6 May 2022. Available at https://www.gov.br/anpd/pt-br/cnpd-2/regimento-interno-cnpd.pdf. Last accessed: 25 July 2022.

_____. Emenda Constitucional nº 115, de 10 de fevereiro de 2022. Altera a Constituição Federal para incluir a proteção de dados pessoais entre os direitos e garantias fundamentais e para fixar a competência privativa da União para legislar sobre proteção e tratamento de dados pessoais. Brasília, 10 Feb. 2022. Available at: http://www.planalto.gov.br/ccivil_03/constituicao/Emendas/Emc/emc115.htm. Last accessed: 27 jun. 2022.

_____. Tribunal Regional do Trabalho da 1ª Região. Acórdão nº 0103519-41.2020.5.01.0000 (MSCiv). Relatora: Desembargadora Raquel de Oliveira Maciel. Rio de Janeiro, 29 April 2021a. Rio de Janeiro, RJ. Available at https://images.jota.info/wp-content/uploads/2021/05/trt1-uber-pericia-algoritmo.pdf. Last Accessed: 1 Feb. 2022.

_____. Autoridade Nacional de Proteção de Dados. Norma de fiscalização da ANPD. Brasília, 28 May 2021b. Available at https://www.gov.br/participamaisbrasil/norma-de-fiscalizacao-da-anpd. Last accessed: 19 July 2022.

_____. Resolução nº 332, de 2020. Dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário e dá outras providências. **Conselho Nacional de Justiça**. Brasil, 21 ago. 2020a.

_____. Projeto de Lei nº 2.630, de 2020. Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet. **Câmara dos Deputados**. Brasil, 2020b.

_____. Projeto de Lei nº 21, de 2020. Estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil; e dá outras providências. **Senado**. Brasil, 2020c.

_____. Mensagem nº 288, de 8 de julho de 2019. **Planalto**. Brasília, DF

_____. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Planalto**. Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Last accessed: 26 June 2022.

_____. Marco Civil da Internet. Planalto, 2014. Available at: http://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Last accessed: 6 June 2022.

_____. Lei de Acesso à Informação. Lei n. 12,527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Planalto, 2011b. Available at: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Last accessed: 6 June 2022.

_____. Lei do Cadastro Positivo. Lei n. 12.414, de 9 de junho de 2011. Disciplina a formação e consulta a bancos de dados com informações de adimplemento, de pessoas naturais ou de pessoas jurídicas, para formação de histórico de crédito. Planalto, 2011a. Available at http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12414.htm. Last accessed 6 June 2022.

_____. Código Civil. Planalto, 2002. Available at: http://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Last accessed on 6 June 2022.

_____, Código de Defesa do Consumidor. Planalto, 1990. Disponível em http://www.planalto.gov.br/ccivil_03/leis/l8078compilado.htm. Last accessed: 6 June 2022.

_____. Constituição de 1988. Available at: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Last accessed: 6 June 2022.

_____. Constituição de 1967. Available at: http://www2.camara.leg.br/legin/fed/consti/1960-1969/constituicao-1967-24-janeiro-1967-365194-publicacaooriginal-1-pl.html. Last accessed: 6 June 2022.

_____. Constituição de 1824. Available at: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao24.htm. Last accessed: 6 June 2022.

BUCHER, Taina. **If... then**: algorithmic power and politics. New York: Oxford University Press, 2018.

BUNDESBEAUFTRAGTE FÜR DATENSCHUTZ UND INFORMATIONSFREIHEIT - BfDI. **Tasks and powers**. BfDI Website, 2019. Available at https://www.bf-

di.bund.de/EN/DerBfDI/UeberUns/DieBehoerde/diebehoerde_node.html. Last accessed 25 July 2022.

BUITEN, Mirian C. **Towards Intelligent Regulation of Artificial Intelligence**. European Journal of Risk Regulation. Vol. 10. 2019, pp. 41–59.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, 1., 2018, S.L. **Proceedings.** S.L.: PMLR, 2018. v. 81, p. 77-91. Available at: http://proceedings.mlr.press/v81/buolamwini18a.html. Last Accessed: 20 jan. 2022.

CAMBRIDGE. Fairness. **Cambridge Dictionary.** Available at https://dictionary.cambridge.org/pt/dicionario/ingles/fairness. Last accessed 12 July 2022.

CATH, Corinne. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. **Philosophical Transactions Of The Royal Society A: Mathematical, Physical and Engineering Sciences**, [S.L.], v. 376, n. 2133, p. 20180080, 15 out. 2018. The Royal Society. http://dx.doi.org/10.1098/rsta.2018.0080.

CENTER FOR AI AND DIGITAL POLICY. **Artificial Intelligence and Democratic Values**: The AI Social Contract Index 2020 (AISCI-2020). Boston, MA, Washington, DC. 2020. Available at: https://dukakis.org/center-for-ai-and-digital-policy/caidp-publishes-artificial-intelligence-and-democratic-values Accessed on 25 November 2021.

CHAPELLE, Olivier; SCHÖLKOPF, Bernhard; ZIEN, Alexander (ed.). **Semi-Supervised Learning**. Cambridge: The Mit Press, 2006.

CHINA. 国家互联网信息办公室关于《互联网信息服务深度合成管理规定（征求意见稿）》公开征求意见的通知. Beijing: Cyberspace Administration of China, 2021. Available at http://www.cac.gov.cn/2022-01/28/c_1644970458520968.htm . Last accessed 28 July 2022.

_____. 互联网信息服务算法推荐管理规定. Beijing, 2022. Available at http://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm. Last accessed: 14 July 2022.

COALIZÃO DIREITOS NA REDE. **CDR solicita ao STF participação em processo que vai definir o envio de recursos para conectividade emergencial na educação básica**. 2021a. Available at https://direitosnarede.org.br/2021/07/28/cdr-solicita-ao-stf-participacao-em-processo-que-vai-definir-o-envio-de-recursos-para-conectividade-emergencial-na-educacao-basica/. Last accessed: 25 July 2022.

_____. **Carta a respeito da composição da lista de expositores das Reuniões Técnicas sobre Relatório de Impacto à Proteção de Dados Pessoais**. 2021b. Available at https://direitosnarede.org.br/2021/06/29/carta-a-respeito-da-composicao-da-lista-de-expositores-das-reunioes-tecnicas-sobre-relatorio-de-impacto-a-protecao-de-dados-pessoais/. Last accessed 25 July 2022.

_____. CDR protocola recurso ao MPF pedindo investigações sobre 'degustação' de dados pessoais. 22 July 2022. Available at https://direitosnarede.org.br/2022/07/21/cdr-protocola-recurso-ao-mpf-pedindo-investigacoes-sobre-degustacao-de-dados-pessoais/. Last accessed: 26 July 2022.

COLOMBO, Cristiano. ANTECEDENTES HISTÓRICOS SOBRE o DIREITO DE PRIVACIDADE NO DIREITO BRASILEIRO. **Direito & TI**, Porto Alegre/RS, 2017. Available at: https://direitoeti.emnuvens.com.br/direitoeti/article/download/74/72. Last accessed: 6 June 2022.

COOLEY, Thomas McIntyre. **A treatise on the law of torts, or, The wrongs which arise independent of contract**. 2. ed. Chicago: Callaghan, 1888.

CORMEN, Thomas H.; LEISERSON, Charles E.; RIVEST, Ronald L.; STEIN, Clifford. **Introduction to Algorithms**. 3. ed. Cambridge: Massachusetts Institute Of Technology, 2009.

COUNCIL OF EUROPE. **ETS nº 108, from 28 January 1981**. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Strasbourg, 1981. Available at: https://www.coe.int/en/web/data-protection/convention108-and-protocol. Last Accessed: 3 Jun. 2022.

COWAN, N. The magical mystery four: How is working memory capacity limited, and why? **Curr. Dir. Psychol. Sci,** 2010, 19, pp. 51–57.

CRAWFORD, Kate. **The Atlas of AI**. New Haven And London: Yale University Press, 2021.

DAVIES, Simon G.. Re-Engineering the Right to Privacy: how privacy has been transformed from a right to a commodity. In: AGRE, Philip E.; ROTENBERG, Marc. **Technology and Privacy**: the new landscape. Cambridge, Massachusetts: The Mit Press, 1997. pp. 143-166.

D'IGNAZIO, Catherine; KLEIN, Laura F. **Data Feminism**. Cambridge, Massachusetts: The MIT Press, 2020.

DAVIES, Harry et al. Uber broke laws, duped police and secretly lobbied governments, leak reveals. **The Guardian**, 11 Jul. 2022. Available at https://www.theguardian.com/news/2022/jul/10/uber-files-leak-reveals-global-lobbying-campaign. Last accessed 21 July 2022.

DÖHMANN, Indra Spiecker. A Proteção De Dados Pessoais Sob O Regulamento Geral De Proteção De Dados Da União Europeia. In: DONEDA, Danilo *et al*. **Tratado de proteção de dados pessoais**. Rio de Janeiro: Forense, 2021. p. 113-129.

DONEDA, Danilo. **Da privacidade à proteção de dados pessoais**: Fundamentos da Lei Geral de Proteção de Dados. 2. ed. São Paulo: Revista dos Tribunais, 2020.

DOSHI-VELEZ, Finale & KIM, Been. Towards a rigorous science of interpretable machine learning. **arXiv***,* 2017, arXiv:1702.08608.

DOSHI-VELEZ, Finale & KIM, Been. Considerations for Evaluation and Generalization in Interpretable Machine Learning. *In* **Explainable and Interpretable Models in Computer Vision and Machine Learning**; Springer, Berlin, Germany, 2018, pp. 3–17.

DOSHI-VELEZ, Finale & KORTZ, Mason. **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.

DOURISH, Paul. Accounting for system behavior: representation, reflection, and resourceful action. In: KYNG, Morten; MATHIASSEN, Lars. **Computers and design in context**. Cambridge, United States: Mit Press, 1997. p. 145-170.

EDWARDS, Lilian & VEALE, Michael. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. **Duke Law & Technology Review,** 4 Dec. 2017, Vol. 16, No. 1, pp. 1-65. Available at: https://ssrn.com/abstract=2972855. Last accessed: 25 Nov. 2021.

EDWARDS, Lilian; VEALE, Michael. Enslaving the Algorithm: from a "right to an explanation" to a "right to better decisions"?. **IEEE Security & Privacy**, [*s. l*], p. 46-54, 2018.

ELECTRONIC FRONTIER FOUNDATION. Face Recognition. **Electronic Frontier Foundation**, 2017. Available at: https://www.eff.org/pages/face-recognition. Last accessed: 19 Jan. 2022.

EUROPEAN COMMISSION. **Liability for Artificial Intelligence and other emerging digital technologies.** 2019. Available at https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en. Last accessed on 25 November 2019.

_____. **Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics.** COM(2020) 64 final. Brussels. 19 February, 2020a.

_____. **White Paper on Artificial Intelligence**: a European approach to excellence and trust. Brussels: European Commission, 2020b.

EUROPEAN DATA PROTECTION BOARD (EDPB). **Guidelines 4/2019 on Article 25 Data Protection by Design and by Default**. Version 2.0, Adopted on 20 October 2020. Brussels: European Commission, 2019.

EUROPEAN DATA PROTECTION SUPERVISOR (EDPS). **EDPS Guidelines on the concepts of controller, processor and joint controllership under Regulation (EU) 2018/1725**. Brussels: European Commission, 2019.

EUROPEAN DIGITAL RIGHTS (EDRi). **The EU AI Act and fundamental rights: Updates on the political process**. EDRi, 2022. Available at https://edri.org/our-work/the-eu-ai-act-and-fundamental-rights-updates-on-the-political-process/. Last accessed 29 July 2022.

EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS (FRA). **Handbook on European data protection law**. Luxembourg: Publications Office of the European Union, 2018.

EUROPEAN UNION. **Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases**. Brussels, 1996.

_____. **Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts**. Brussels, 2020.

EUROPEAN PARLIAMENT. **European Parliament legislative resolution of 5 July 2022 on the 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.** Brussels, 2022.

FJELD, Jessica *et al*. Principled Artificial Intelligence: mapping consensus in ethical and rights-based approaches to principles for ai. **Berkman Klein Center For Internet & Society**, 2020.

FRAGOSO, Nathalie. O impacto do Marco Civil sobre a proteção da privacidade no Brasil. Especial Marco Civil 5 Anos. InternetLab, 2019. Available at https://internetlab.org.br/pt/especial/o-impacto-do-marco-civil-sobre-a-protecao-da-privacidade-no-brasil/. Last accessed: 6 June 2022.

GARCIA, Renata Cavalcanti de Carvalho. **Proteção de dados pessoais no Brasil:** Uma análise da Lei nº 13.709/2018 sob a perspectiva da Teoria da Regulação Responsiva. Journal of Law and Regulation, *[S. l.]*, v. 6, n. 2, p. 45–58, 2020. Disponível em: https://periodicos.unb.br/index.php/rdsr/article/view/28490. Acesso em: 4 abr. 2021.

GEREKE, Marika; BRÜHL, Tanja. Unpacking the unequal representation of Northern and Southern NGOs in international climate change politics. **Third World Quarterly**, [S.L.], v. 40, n. 5, p. 870-889, 12 abr. 2019. Informa UK Limited. http://dx.doi.org/10.1080/01436597.2019.1596023.

GERMANY. Bundesdatenschutzgesetz vom 30. Juni 2017 (BGBl. I S. 2097). Available at https://www.gesetze-im-internet.de/bdsg_2018/BJNR209710017.html#BJNR209710017BJNG000100000. Last accessed: 25 July 2022.

_____. BVerfG. Urteil Des Ersten Senats Vom 15. Dezember 1983 no. 1 BvR 209/83 -, Rn. 1-215. **Germany**, 1983.

GOODMAN, Bryce; FLAXMAN, Seth. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation'. **ICML Workshop on Human Interpretability in Machine Learning**, 2016. arXiv:1606.08813 (v3).

GOOGLE DEVELOPERS. **Introduction to Machine Learning Problem Framing**: Common ML Problems. Available at: https://developers.google.com/machine-learning/problem-framing/cases#youtube-watch-next. Last accessed: 14 jan. 2022.

GOMES, Luciano. **Faraó** (Divindade do Egito). In MENEZES, Margareth. Brasileira ao Vivo: Uma Homenagem ao Samba-Reggae. EMI: 2006. 03m19s.

GONÇALVES, M. E. **The risk-based approach under the new EU data protection regulation:** a critical perspective. Journal of Risk Research, 2019, DOI: 10.1080/13669877.2018.1517381.

GRAY, Mary L.; SURI, Siddharth. **Ghost Work**: how to stop Silicon Valley from building a new global underclass. New York: Houghton Mifflin Harcourt, 2019.

GRÖGER, Christoph. **There Is No AI Without Data**. Communications of the ACM, November 2021, Vol. 64 No. 11, Pages 98-108 10.1145/3448247 Available at https://cacm.acm.org/magazines/2021/11/256400-there-is-no-ai-without-data/fulltext. Accessed on 22 November 2021.

GURKAYNAK, Gonenc; YILMAZ, Ilay; HAKSEVER, Gunes. Stifling artificial intelligence: human perils. **Computer Law & Security Review**, [S.L.], v. 32, n. 5, p. 749-758, out. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.clsr.2016.05.003.

HALEVY, Alon.; NORVIG, Peter.; PEREIRA, Fernando. **The Unreasonable Effectiveness of Data**. IEEE Intelligent Systems. Los Alamitos, CA: IEEE Computer Society, 2009 no. 24, p. 8-12. Available at: https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/35179.pdf. Last accessed: 19 Jan. 2022.

HARA, Kotaro *et. al*. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In: CHI '18, 4., 2018, Montreal. **Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.** Montreal: Research Collection School Of Information Systems, 2018. p. 1-14. Available at: https://ink.library.smu.edu.sg/sis_research/4209. Last accessed: 03 Feb. 2022.

HAMON, Ronan *et al*. Bridging the Gap Between AI and Explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. **IEEE Computational Intelligence Magazine,** pp. 72-85. Feb. 2022.

HEAVEN, Will Douglas. Bias isn't the only problem with credit scores—and no, AI can't help. **MIT Technology Review**. 17 June 2021. Available at https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/. Accessed on 18 November 2021.

HIGH LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (HLEG). **A definition of AI**: main capabilities and scientific disciplines. Brussels: European Commission, 2019.

HIGH LEVEL EXPERT GROUP FOR ARTIFICIAL INTELLIGENCE (HLEG). **Ethics guidelines for trustworthy AI**. Brussels: European Commission, 2019.

HILDEBRANDT, Mireille. Slaves to Big Data. Or Are We? **IDP.** Journal promoted by the Law and Political Science Department. Barcelona: Universitat Oberta de Catalunya, 2013, No. 17, pp. 27-44. Available at http://journals.uoc.edu/index.php/idp/article/view/n17-hildebrandt/n17-hildebrandt-en http://dx.doi.org/10.7238/idp.v0i17.1977. Last accessed: 19 jan. 2022.

HILL, Kashmir. Wrongfully Accused by an Algorithm. **New York Times.** New York. 24 jul. 2020. Available at: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html. Last accessed: 20 jan. 2022.

HOWARD, Philip N. **Lie Machines**: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives. New Haven, London: Yale University Press, 1st Edition, 2020.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS (IEEE). Ethically Aligned Design, First Edition, 2021. Available at: https://ethicsinaction.ieee.org/#ead1e. Last accessed: 17 May 2022.

GARCIA, Renata Cavalcanti de Carvalho. **Proteção de dados pessoais no Brasil:** Uma análise da Lei nº 13.709/2018 sob a perspectiva da Teoria da Regulação Responsiva. Journal of Law and Regulation, *[S. l.]*, v. 6, n. 2, p. 45–58, 2020. Disponível em: https://periodicos.unb.br/index.php/rdsr/article/view/28490. Acesso em: 4 abr. 2021.

INSTITUTO BRASILEIRO DE DEFESA DO CONSUMIDOR (IDEC). **Justiça suspende sistema de reconhecimento facial no Metrô de SP. 2022.** Available at https://idec.org.br/noticia/justica-determina-suspensao-de-sistema-de-reconhecimento-facial-no-metro-de-sp. Last accessed 28 July 2022.

IRAMINA, A. GDPR v. GDPL: Strategic Adoption of the responsiveness approach in the elaboration of Brazil´'s General Data Protection Law and the EU General Data Protection Regulation**. Law, State and Telecommunications Review**, [S. l.], v. 12, n. 2, p. 91–117, 2020. DOI: 10.26512/lstr.v12i2.34692. Disponível em: https://periodicos.unb.br/index.php/RDET/article/view/34692. Acesso em: 13 may. 2021

IRANI, Lilly. The hidden faces of automation. **XRDS**: Crossroads, The ACM Magazine for Students, [S.L.], v. 23, n. 2, p. 34-37, 15 dez. 2016. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/3014390. Available at: https://dl.acm.org/doi/fullHtml/10.1145/3014390. Last accessed: 31 mar. 2022.

KAMINSKI, Margot E. The Right to Explanation, Explained. **U of Colorado Law Legal Studies Research Paper**, n. 18-24, 15 June 2018.

KAMINSKY, Margot E. & MALGIERI, Gianclaudio. Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. International Data Privacy Law, **U of Colorado Law Legal Studies Research Paper** No. 19-28, 2020. Available at: https://ssrn.com/abstract=3456224.

KANDPAL, M., KRISHNAN, P., & SAMAVEDHAM, L., **Data driven fault detection using multi-block PLS based path modeling approach**. 2012. 11th International Symposium on Process Systems Engineering, 1291–1295. doi:10.1016/b978-0-444-59506-5.50089-4.

KHANBHAI, Mustafa et al.. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. **BMJ Health & Care Informatics**, [S.L.], v. 28, n. 1, Mar. 2021. Available at http://dx.doi.org/10.1136/bmjhci-2020-100262. Last accessed: 15 July 2022.

KNOTH, Pedro. Receita autoriza Serpro a compartilhar CPF, telefone e mais com terceiros. Tecnoblog. 19 April 2022. Available at: https://tecnoblog.net/noticias/2022/04/19/receita-autoriza-serpro-a-compartilhar-cpf-telefone-e-mais-com-terceiros/. Last accessed: 6 June 2022.

KOLIEB, Jonathan. When to Punish, When to Persuade and When to Reward: Strengthening Responsive Regulation with the Regulatory Diamond. **Monash University Law Review**, v. 41, n. 1, p. 136-162, 2015.

KUNIAVSKY, M. **Smart Things**: Ubiquitous Computing User Experience Design. 2010. doi:10.1016/b978-0-12-374899-7.00001-1.

LABORATÓRIO DE POLÍTICAS PÚBLICAS E INTERNET - LAPIN. **Atuação do LAPIN como amicus curiae na ADI 6387**. 2020. Available at https://lapin.org.br/2020/05/28/atuacao-do-lapin-como-amicus-curiae-na-adi-6387/. Last accessed: 25 July 2021.

LAMBERT, Fred. Breakdown of raw materials in Tesla's batteries and possible bottlenecks. **Electrek.** Available at: https://electrek.co/2016/11/01/breakdown-raw-materials-tesla-batteries-possible-bottleneck/. Last accessed: 02 feb. 2022.

LARSON, J. *et al*. (2016). How We Analyzed the COMPAS Recidivism Algorithm. **Pro Publica.** Available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism- algorithm. Last accessed: 16 Dec. 2021.

LEMOS, Alessandra et al. **Nota Técnica PL n. 21/2020**: sobre o marco legal do desenvolvimento e uso da inteligência artificial no brasil. LAPIN, Brasília, v. 1, n. 1, p. 1-49, nov. 2021. Available at: https://lapin.org.br/2021/11/09/nota-tecnica-atualizada-discute-o-pl-21-a-2020-do-marco-legal-de-ia/. Last access: 12 jul. 2022.

LEMOS, Ronaldo *et al*. Brazilian General Data Protection Law (LGPD, English translation). **IAPP**, [*s. l*], Oct. 2020. Available at: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/. Last accessed: 14 June, 2022.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. **Entropy,** 2021, 23, 18. Available at: https://dx.doi.org/10.3390/e23010018. Last accessed: 19 May 2022.

LIPTON, Zachary C. The Mythos of Model Interpretability. **ICML Workshop on Human Interpretability in Machine Learning** (WHI). New York, NY, USA, 2016, pp. 96-100.

MALGIERI, Gianclaudio. **Automated Decision-Making in the EU Member States**: The Right to Explanation and Other 'Suitable Safeguards' for Algorithmic Decisions in the EU National Legislations.. Computer Law & Security Review 2019 Forthcoming. 17 August 2018. Available at SSRN: https://ssrn.com/abstract=3233611 or http://dx.doi.org/10.2139/ssrn.3233611. Accessed on 25 November 2021.

MALGIERI, Gianclaudio. "Just" Algorithms: AI Justification (beyond explanation) in the GDPR. **Gianclaudio Malgieri,** 14 Dec. 2020. Available at https://www.gianclaudiomalgieri.eu/2020/12/14/just-algorithms/. Last accessed: 29 jan. 2022.

MANOVICH, Lev. Trending: the promises and the challenges of big social data. In: GOLD, Matthew K. (ed.). **Debates in the Digital Humanities**. Minneapolis: The University Of Minnesota Press, 2011. p. 460-475. Available at: http://www.arise.-mae.usp.br/wp-content/uploads/2018/03/Debates-in-the-digital-humanities-1.pdf. Last accessed: 19 jan. 2022.

MARDA, Vidushi. Machine Learning and Transparency: A scoping exercise, 2018.

MAYER-SCHÖNBERGER, Viktor. Generational Development of Data Protection in Europe. In: AGRE, Philip E.; ROTENBERG, Marc. **Technology and Privacy**: the new landscape. Cambridge, Massachusetts: The Mit Press, 1997. pp. 219-241.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big Data**: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt, 2013.

MCGOEY, Linsey. The logic of strategic ignorance. **The British Journal Of Sociology**, [*s. l*], v. 63, n. 3, p. 533-576, Sep. 2012.

MCGOUGH, M. How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say. **Sacramento Bee**. 7 Aug. 2018.

MENDES, Laura Schertel. **Privacidade, proteção de dados e defesa do consumidor**: linhas gerais de um novo direito fundamental. São Paulo: Saraiva, 2014.

MENDES, Laura Schertel; DONEDA, Danilo. Reflexões iniciais sobre a nova Lei Geral de Proteção de Dados. Revista de Direito do Consumidor, S.L., v. 120, p. 469-483, nov. 2018.

MENDES, Laura Schertel; FONSECA, Gabriel Campos Soares da. STF reconhece direito fundamental à proteção de dados. **Revista dos Tribunais**, 30 jun. 2020. Available at: https://www.researchgate.net/profile/Gabriel-Campos-Soares-Da-Fonseca/publication/344381892_STF_reconhece_direito_fundamental_a_protecao_de_dados/links/5f6e79fe92851c14bc97260e/STF-reconhece-direito-fundamental-a-protecao-de-dados.pdf. Last accessed: 6 June 2022.

MENDES, Laura Schertel et al. Comparative Analysis between Personal Data Legal Protection: A Methodological Approach. **CPRLATAM Conference**, Cordoba, Argentina, 1-2 July, 2019 in coordination with CLT2019, 1-5 July, 2019.

MENKE, Fabiano; GOULART, Guilherme Damásio. Segurança da Informação e Vazamento de Dados. In: DONEDA, Danilo et al. **Tratado de proteção de dados pessoais**. Rio de Janeiro: Forense, 2021. p. 350-370.

MERLER, Michele *et al*. Diversity in Faces. **Arvix**, [*s. l*], jan. 2019. Available at: https://arxiv.org/abs/1901.10436. Last accessed: 20 Jan. 2022.

META. **How advertisers' audience selections appear in "Why am I seeing this ad?"**. Available at: https://www.facebook.com/business/m/why-am-i-seeing-this-ad. Last accessed: 25 Feb. 2022.

MILLER, Tim. Explanation in artificial intelligence: insights from the social sciences. **Artificial Intelligence.** S.L., pp. 1-38, 2019.

MIRAGEM, Bruno. A lei geral de proteção de dados (Lei 13.709/2018) e o direito do consumidor. Revista dos Tribunais, v. 1009, p. 173-222, nov., 2019. p. 2.

MITTELSTADT, Brent; RUSSEL Chris & WACHTER, Sandra. **Explaining Explanations in AI**. FAT*'19. January 2019. Atlanta, USA. Available at https://doi.org/10.1145/3287560.3287574. Last accessed on 25 November 2021.

MITTELSTADT, Brent. Principles alone cannot guarantee ethical AI. **Nature Machine Intelligence**, Volume 1, pp. 501–507 (2019). Available at https://www.nature.com/articles/s42256-019-0114-4. Last accessed: 16 May 2022.

MOHSENI, Sina; RAGAN, Eric; HU, Xia. Open Issues in Combating Fake News: interpretability as an opportunity. **Arvix**, [*s. l*], 04 apr. 2019.

MOHSENI, Sina; ZAREI, Niloofar; RAGAN, Eric D.. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. **Arvix**, [*s. l*], 28 nov. 2018.

MOLNAR, Christoph. **Interpretable Machine Learning**: A Guide for Making Black Box Models Explainable. 2nd Edition, 2022. Available at: https://christophm.github.io/interpretable-ml-book/. Last accessed: 20 May 2022.

MONTEIRO, Renato Leite. Existe um direito à explicação na Lei Geral de Proteção de Dados Pessoais do Brasil? **Artigo Estratégico 24**. Rio de Janeiro: Instituto Igarapé, Dec 2018.

MORAES, Thiago Guimarães Moraes et al. Open data and the COVID-19 pandemic: anonymisation as a technical solution for transparency, privacy, and data

protection. **International Data Privacy Law**, Volume 11, Issue 1, February 2021, pp. 32–47. Available at https://doi.org/10.1093/idpl/ipaa025. Last accessed: 6 June 2022.

MORAES, Thiago Guimarães; ALMEIDA, Eduarda Costa; PEREIRA, José Renato Laranjeira de. Smile, you are being identified!: risks and measures for the use of facial recognition in (semi-)public spaces. **AI Ethics**, [*s. l*], v. 1, p. 159-172, May 2021.

MORESCHI, Bruno *et al*. The Brazilian Workers in Amazon Mechanical Turk: dreams and realities of ghost workers. **Contracampo – Brazilian Journal Of Communication**, [*s. l*], v. 39, n. 1, p. 44-64, mar. 2020.

NEWTON, Casey. Half of all Facebook moderators may develop mental health issues. **The Verge.** 13 May 2020. Available at: https://www.theverge.com/interface/2020/5/13/21255994/facebook-content-moderator-lawsuit-settlement-mental-health-issues. Last Accessed: 02 Feb. 2022.

NG, Andrew. **Machine Learning**: unsupervised learning. Unsupervised Learning. Coursera. Available at: https://www.coursera.org/learn/machine-learning/lecture/olR-Zo/unsupervised-learning. Last accessed: 14 jan. 2022.

NILSSON, Nils J.. **The Quest for Artificial Intelligence**: a history of ideas and achievements. Cambridge: Cambridge University Press, 2009. Available at: https://ai.stanford.edu/~nilsson/QAI/qai.pdf. Last accessed: 19 jan. 2022.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). **Face Recognition Vendor Test (FRVT) Part 3**: Demographic Effects. 2019. Available at https://doi.org/10.6028/NIST.IR.8280 (2019). Accessed on 20 Jan 2020.

OBERMEYER, Z. & MULLAINTHAN, S. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70M People. **Proceedings of the Conference on Fairness, Accountability, and Transparency** [online], 2019. Available at https://dl.acm.org/doi/10.1145/3287560.3287593. Accessed on 10 April 2020.

OECD. Recommendation nº OECD/LEGAL/0463, from 23 September 1980. **OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data**, 1980.

OECD. **OECD Framework for the Classification of AI Systems**: a tool for effective AI policies. Paris: OECD Digital Economy Papers, OECD Publishing, 2022. Available at: https://doi.org/10.1787/cb6d9eca-en. Last accessed: 17 May 2022.

ONETRUST. Brazil: Idec requests investigation into leakage of 16 million Brazilian COVID-19 patients' data. Data Guidance. 30 Nov. 2020. Available at: https://www.-dataguidance.com/news/brazil-idec-requests-investigation-leakage-16-million. Last accessed: 6 June 2022.

PASQUALE, Frank. **The Black Box Society**: the secret algorithms that control money and information. Cambridge, Massachusetts: Harvard University Press, 2015.

PEREIRA, José Renato Laranjeira de; MORAES, Thiago Guimarães. Promoting irresponsible AI: lessons from a Brazilian bill. **Heinrich Böll Stiftung Blog,** Brussels, 14 Feb. 2022a. Available at https://eu.boell.org/en/2022/02/14/promoting-irresponsible-ai-lessons-brazilian-bill. Last access: 12 Jul. 2022.

_____. Data-hungry government in Brazil: how narratives about state efficiency became fuel for personal data sharing. **Heinrich Böll Stiftung Blog**, 7 June 2022b. Available at: https://eu.boell.org/en/2022/06/07/data-hungry-government-brazil. Last accessed: 27 June 2022.

PIKETTY, Thomas. Capital and Ideology. Translated by Arthur Goldhammer, Cambridge: The Belknap Press of Harvard University Press, 2020.

PORTUGAL, Ivens; ALENCAR, Paulo; COWAN, Donald. The use of machine learning algorithms in recommender systems: a systematic review. **Expert Systems With Applications,** [s. l], v. 97, p. 205-227, 1 May 2018.

RADER, Emilee *et al*. Explanations as Mechanisms for Supporting Algorithmic Transparency. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2018, Montréal, Canada. **Paper.** Montréal, Canada: Chi, 2018.

REIS, Carolina; ALMEIDA, Eduarda; DA SILVA; Felipe; DOURADO, Fernando. **Relatório sobre o uso de tecnologias de reconhecimento facial e câmeras de vigilância pela administração pública no Brasil**. Brasília: Laboratório de Políticas Pú-

blicas e Internet, 2021. Available at: https://lapin.org.br/2021/07/07/vigilancia-automatizada-uso-de-reconhecimento-facial-pela-administracao-publica-no-brasil/. Accessed on 18 November 2021.

REMIX. Italian city plans controversial social credit system. **REMIX**, 27 Apr. 2022. Available at https://rmx.news/italy/italian-city-plans-controversial-social-credit-system/. Last accessed 1 Aug. 2022.

ROBBINS, Scott. A Misdirected Principle with a Catch: explicability for AI. **Minds And Machines**, [*s. l*], v. 29, p. 495-514, 2019.

RODOTÀ, Stefano. *Tecnologie e diritti*, Bologna: Il Mulino, 1995.

ROTH, Emma. Facebook content moderators protest low wages with mobile billboard 4 Moderators employed by Accenture will run a mobile billboard to demand higher pay. **The Verge.** 19 out. 2021. Available at: https://www.theverge.com/interface/2020/5/13/21255994/facebook-content-moderator-lawsuit-settlement-mental-health-issues. Last accessed: 02 Feb. 2022.

RUDIN, C. **Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead**. 2019. Available at: https://arxiv.org/abs/1811.10154. Last accessed 28 Aug 2020.

RUSSEL, Peter; NORVIG, Stuart. **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SADOWSKI, Jathan. Potemkin AI. **Real Life Mag.** 06 Aug. 2018. Available at: https://reallifemag.com/potemkin-ai/. Last accessed: 02 Feb. 2022.

SENADO. Comissão de juristas da inteligência artificial faz balanço de audiências públicas. Brasília: **Agência Senado**, Brasília, 16 May 2022a. Available at https://www12.senado.leg.br/noticias/materias/2022/05/16/comissao-de-juristas-da-inteligencia-artificial-faz-balanco-de-audiencias-publicas. Last access: 12 July 2022.

_____. MP concede autonomia de autarquia à Autoridade Nacional de Proteção de Dados. Brasília: **Agência Senado,** 14 June 2022b. Available at https://www12.sena-

do.leg.br/noticias/materias/2022/06/14/mp-concede-autonomia-de-autarquia-a-autoridade-nacional-de-protecao-de-dados. Last accessed 29 July 2022.

SÉNIT, C.-A.; BIERMANN, F. In Whose Name Are You Speaking? The Marginalization of the Poor in Global Civil Society. **Glob. Policy**, 12, 2021, pp. 581-591. Available at https://doi.org/10.1111/1758-5899.12997. Last accessed 25 July 2022.

SHALEV-SCHWARTZ, Shai; BEN-DAVID, Shai. **Understanding Machine Learning**: from theory to algorithms. New York: Cambridge University Press, 2014.

SHORTLIFFE, Edward. **Computer-Based Medical Consulations**: MYCIN. Elsevier, 1976.

SILVA, Tarcízio. **Racismo algorítmico:** inteligência artificial e discriminação nas redes digitais. São Paulo, Edições Sesc, 2022.

SOUZA, Carlos Affonso, PERRONE, Christian, MAGRANI, Eduardo, "O direito à explicação entre a experiência europeia e a sua positivação na LGPD" in DONEDA, Danilo et al., **Tratado de proteção de dados pessoais**, Rio de Janeiro, Forense, 2021, p. 243-270.

STATEWATCH. EU has spent over €340 million on border AI technology that new law fails to regulate. **Statewatch**, 12 May 2022. Available at: https://www.statewatch.org/news/2022/may/eu-has-spent-over-340-million-on-border-ai-technology-that-new-law-fails-to-regulate/. Last accessed 1 Aug. 2022.

STOHL, C.; STOHL M.; LEONARDI P. M. Managing opacity: information visibility and the paradox of transparency in the digital age. **International Journal of Communication Systems**, 2016, Vol 10, pp. 123–137.

TAYLOR, Linnet. Safety in Numbers?: group privacy and big data analytics in the developing world. In: TAYLOR, Linnet; FLORIDI, Luciano; SLOOT, Bart van Der. **Group Privacy**: new challenges of data technologies. Cham: Springer, 2017. Cap. 2. p. 13-36.

TOSONI, Luca; BYGRAVE, Lee A.. Article 4: definitions. In: KUNER, Christopher; BYGRAVE, Lee A.; DRECHSLER, Laura. The EU General Data Protection Regulation (GDPR): a commentary. Oxford: **Oxford University Press**, 2020. p. 103-115.

TRETYAKOV, Konstantin. Machine learning techniques in spam filtering. **Data Mining Problem-Oriented Seminar, MTAT**, S. L., v. 3, n. 177, p. 60-79, May 2004.

TRIBUNAL DE CONTAS DA UNIÃO (TCU). Acórdão No. 1970/2017, j. 06/09/2017, plenário. **Processo no 029.688/2016-7**. Relator Aroldo Cedraz.

UNITED NATIONS. Report Of The Special Rapporteur On The Promotion And Protection Of The Right To Freedom Of Opinion And Expression nº A/73/348, de 2018. Transmits report of the Special Rapporteur, David Kaye, submitted in accordance with Human Rights Council resolution 34/18. **Secretary-General of the United Nations**. Available at: https://digitallibrary.un.org/record/1643488?ln=en. Last accessed: 31 May 2022.

UNITED STATES OF AMERICA. The Computer and Invasion of Privacy. Hearings before a subcommittee of the committee on government operations house of representatives. July 26, 27 and 28, 1966. p. 7. **US House of Representatives,** 1966.

_____. H.R.6580 Algorithmic Accountability Act of 2022. Washington, 2022. Available at https://www.congress.gov/bill/117th-congress/house-bill/6580/text. Last Accessed 14 July 2022.

URUPÁ, Marcos. Regulação responsiva dará o tom da agenda regulatória da ANPD nos próximos dois anos. **Teletime**, 28 Jan 2021. Available at: https://teletime.com.br/28/01/2021/regulacao-responsiva-dara-o-tom-da-agenda-regulatoria-da-anpd-nos-proximos-dois-anos/. Last accessed 19 July 2022.

VATICAN. **Rome Call for AI Ethics**. Rome, 28 Feb 2020. Available at https://www.vatican.va/roman_curia/pontifical_academies/acdlife/documents/rc_pont-acd_life_doc_20202228_rome-call-for-ai-ethics_en.pdf. Last accessed 14 July 2022.

VEALE, Michael. **Governing Machine Learning that Matters**. 2019. 352 pp. Doctoral Thesis. Doctor Of Philosophy In Science, Technology, Engineering And Public Policy, University College London, London, 2019.

WACHTER, Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. **IDPL**, n. 76, 2017.

WACHTER, Sandra; MITTELSTADT, Brent and RUSSEL Chris. Counterfactual Explanations Without Opening The Black Box: Automated Decisions and the GDPR. **Harvard Journal Of Law & Technology**, [*s. l*], v. 31, n. 2, p. 841-887, Spring 2018.

WACHTER, Sandra; MITTELSTADT, Brent and RUSSEL Chris. **Explaining Explanations in AI**. FAT*'19. January 2019. Atlanta, USA. Available at https://doi.org/10.1145/3287560.3287574. Last accessed on 25 November 2021.

WALL STREET JOURNAL. the facebook files. **Wall Street Journal**, 2021. Available at https://www.wsj.com/articles/the-facebook-files-11631713039. Last accessed 21 July 2022.

WERNECK, Antônio. Reconhecimento facial falha em segundo dia, e mulher inocente é confundida com criminosa já presa. **O Globo.** Rio de Janeiro, p. 1-2. 11 jul. 2019. Available at: https://oglobo.globo.com/rio/reconhecimento-facial-falha-em-segundo-dia-mulher-inocente-confundida-com-criminosa-ja-presa-23798913. Last accessed: 20 jan. 2022.

WEXLER, R. When a Computer Program Keeps You in Jail: How Computers are Harming Criminal Justice". **New York Times**. 13 June 2017.

WILSON, B., HOFFMAN, J. et al Predictive Inequity in Object Detection. **Arxiv.** Available at https://arxiv.org/pdf/1902.11097.pdf. Last accessed 10 April 2020.

WIMMER, Miriam; DONEDA, Danilo. "Falhas de IA" e a Intervenção Humana em Decisões Automatizadas: Parâmetros para a Legitimação pela Humanização. **Revista Direito Público**, Brasília, Volume 18, n. 100, 374-406, out./dez. 2021.

YIN, Ming *et al*. The Communication Network Within the Crowd. **Proceedings Of The 25Th International Conference On World Wide Web**, Montreal, p. 1293-1303, 11 Apr. 2016. http://dx.doi.org/10.1145/2872427.2883036.

ZUBOFF, Shoshana. **The Age of Surveillance Capitalism**: the fight for a human future at the new frontier of power. New York: Public Affairs, 2019.