Transparência pela cooperação: como a regulação responsiva pode auxiliar na promoção de sistemas de *machine-learning* inteligíveis

Transparency through cooperation: How responsive regulation may help promote intelligible machine learning systems

Submetido(*submitted*): 13/05/2021
Parecer(*revised*): 20/05/2021
Aceito(*accepted*): 27/05/2021

Artigo submetido à revisão cega por pares (*Article submitted to peer blind review*)

José Renato Laranjeira de Pereira*
https://orcid.org/0000-0002-9605-8121

Abstract

[Purpose] *To analyze the applicability of the theory of responsive regulation to promote the intelligibility of machine learning systems under the focus of the Brazilian General Data Protection Law (LGPD).*

[Methodology/approach/design] *This article has the theory of responsive regulation as a theoretical framework and will initially be based on a comparative analysis of the LGPD and the GDPR to identify how this theory can assist Brazilian regulators, more specifically the National Data Protection Authority, to address intelligibility of artificial intelligence systems.*

[Findings] *From a comparative analysis between how LGPD and GDPR deal with the issue of automated decision systems (including machine-learning) explainability, this article identified that the rationale for cooperation between regulator and regulated, a network governance system and the existence of a regulatory pyramid allows for the application of the theory of responsive regulation to promote the intelligibility of these systems.*

[Practical implications] *AI systems have often been accused of discriminatory bias, something which may increase the racial and gender gaps in Brazil. Ensuring that the technology is understandable for humans to better identify how to address these shortcomings is paramount to promoting the use of fairer systems. This study intends, by identifying the most appropriate regulatory strategies to deal with algorithmic opacity, to assist regulators in addressing the discrimination promoted by these systems.*

Keywords*: Responsive regulation. Artificial intelligence. Bias. Intelligibility. Transparency.*

*Diretor do Laboratório de Políticas Públicas e Internet - LAPIN, Mestrando em Direito Regulatório pela Universidade de Brasília - UnB e Bacharel em Direito pela UnB com intercâmbio na Università degli Studi di Roma Tre. É German Chancellor Fellow da turma 2021-2022 pela Fundação Alexander von Humboldt. E-mail: joserenato@lapin.org.br.

Resumo

[Propósito] Analisar a aplicabilidade da teoria da regulação responsiva para promoção da inteligibilidade de sistemas de *machinelearning* sob o enfoque da Lei Geral de Proteção de Dados.

[Metodologia/abordagem/design] Este artigo tem a teoria da regulação responsiva como marco teórico e se baseará, inicialmente, em uma análise comparada da LGPD e do RGPD para identificar como essa teoria pode auxiliar reguladores brasileiros, mais especificamente a Autoridade Nacional de Proteção de Dados, a abordar a inteligibilidade de sistemas de inteligência artificial.

[Resultados] A partir de uma análise comparativa prévia entre como LGPD e RGPD lidam com o tema da explicabilidade de sistemas de decisão automatizada (incluso *machine-learning*), identificou-se que o racional de cooperação entre regulador e regulado, um sistema de governança em rede e a existência de uma pirâmide regulatória permitem a aplicação da teoria da regulação responsiva para a promoção da inteligibilidade desses sistemas.

[Implicações práticas] Sistemas de IA têm sido frequentemente acusados de possuírem vieses discriminatórios. Isso faz com que pessoas negras sejam mais frequentemente identificadas do que brancas por tecnologias de reconhecimento facial ou que afrodescendentes tenham menor chance de conseguir crédito, potencializando o abismo racial no Brasil. Garantir que a tecnologia seja compreensível para humanos identificarem melhor como endereçar essas falhas é primordial para promover o uso de sistemas mais justos. O presente estudo pretende, por meio da identificação das estratégias regulatórias mais adequadas a lidar com a opacidade algorítmica, auxiliar reguladores a endereçar a discriminação promovida por esses sistemas.

Palavras-chave: Regulação responsiva. Inteligência artificial. Viés discriminatório. Inteligibilidade. Transparência.

INTRODUÇÃO

Sistemas de inteligência artificial dominam múltiplos aspectos da experiência humana. Ubíquos, estão presentes em ambientes como filtros de *spam*, redes sociais, sistemas de busca, *smart TVs* e câmeras de vigilância. Apesar desse caráter onipresente, a complexidade de seu funcionamento faz com que poucos desses sistemas sejam compreensíveis por humanos, o que dificulta compreender o real impacto que geram na sociedade e como responsabilizar aqueles que os desenvolvem ou utilizam.

Para endereçar isso, reguladores devem adotar estratégias regulatórias que levem em conta as densas assimetrias de informação existentes entre diferentes atores na cadeia de produção e aplicação da tecnologia, de modo a incentivar a adoção de sistemas inteligíveis, seja por serem interpretáveis ou explicáveis.

Com isso em vista, o presente artigo analisa a aplicabilidade da teoria da regulação responsiva, marcada por um caráter dialógico fundado em incentivos intrínsecos e extrínsecos visando buscar o máximo de cooperação entre reguladora e regulada, para conceber estratégias regulatórias adequadas a lidar com a frequente opacidade de sistemas de inteligência artificial.

Inicialmente, conceituaremos o que são sistemas de aprendizagem de máquina (*machine-learning*) e de que forma eles podem impactar o exercício de direitos e liberdades de indivíduos e grupos, principalmente marginalizados (Seção II). Em seguida, nos debruçaremos sobre o que significa tornar esses sistemas inteligíveis (Seção III). A partir disso, vamos analisar como a Lei Geral de Proteção de Dados (LGPD) e o Regulamento Geral de Proteção de Dados europeu (RGPD) lidam com o tema da explicabilidade de sistemas de *machine-learning* (Seção IV), bem como o que significam abordagens baseadas em risco (Seção V). A partir disso, passamos a uma reflexão sobre quais informações devem ser providas por uma tecnologia dessa natureza (Seção VI)

Finalmente, analisaremos de que forma a teoria da regulação responsiva pode ser útil para a promoção de maior transparência desses sistemas, de modo a promover a proteção de direitos de indivíduos e grupos.

SISTEMAS DE APRENDIZADO DE MÁQUINA

De acordo com Shalev-Shwartz e Ben-David, o aprendizado de máquina, ou *machine-learning* (ML), refere-se à detecção automatizada de padrões significativos em dados e se tornou uma tecnologia comum aplicada para tarefas que exigem extração de informações de grandes conjuntos de dados - intimamente relacionado ao que é chamado de *big data*. Esses sistemas, aplicados em espaços que vão desde motores de busca e *feeds* de mídia social para sistemas de pontuação de crédito e aplicativos de justiça preditiva, são um ramo da inteligência artificial e se tornaram quase onipresentes (SHALEV-SHWARTZ e BEN-DAVID, 2014).

Os sistemas de ML "aprendem" a realizar atividades específicas, extraindo informações dos dados de treinamento. Ao aprenderem a partir de grandes quantidades de informações, eles processam os dados recebidos (entrada, *input*) para realizar uma tarefa, geralmente uma previsão baseada na experiência passada (saída, *output*). Já o conjunto finito de regras que descreve uma sequência de operações a serem seguidas pelo sistema para a solução de um problema específico consiste no chamado algoritmo (JOLLIFFE, 2011).

A maioria dos sistemas de ML foram descritos como caixas pretas, no sentido de que frequentemente não são capazes de explicar suas previsões de uma forma que os humanos possam entender. Essa falta de transparência é seguida pela falta de responsabilização de quem faz uso desses sistemas

(MOHSENI, RAGAN, 2018). Afinal, sem serem capazes de compreendê-los, juízes e reguladores não podem abordar adequadamente quem deve ser responsável por um resultado negativo feito por um sistema.

Alguns exemplos de consequências graves de sistemas de ML referem-se a sistemas de saída enviesados que negam liberdade condicional para presidiários (WEXLER, 2017) e modelos de poluição baseados em ML que afirmaram erroneamente que o ar altamente poluído de uma cidade era seguro para respirar (MCGOUGH, 2018). No entanto, fornecer transparência em sistemas de ML também representa um desafio de equilibrar as proteções de sigilo comercial detidas por empresas que implantam ou fazem uso desses sistemas, um recurso que impõe mais barreiras para sua inteligibilidade.

Bucher (2018) argumenta que definir algoritmos como caixas pretas pode, às vezes, consistir em uma espécie de "desconhecido estratégico" (*strategic unknown*), em que organizações e indivíduos têm confortavelmente evitado muitos esforços para tornar seus sistemas inteligíveis a fim de evitar responsabilidades. Tornar esses sistemas explicáveis foi, portanto, considerado um instrumento importante para permitir que reguladores e usuários entendessem melhor como esses modelos de tomada de decisão automatizados alcançam previsões específicas e como revisá-los em caso de erros, permitindo que desenvolvedores e usuários tenham maior controle sobre esses sistemas.

No entanto, Arya et al. (2019) identificaram que há uma lacuna entre o que a comunidade técnica está produzindo sobre explicabilidade e o que os reguladores e a sociedade como um todo exigem desses sistemas. Uma razão para essa lacuna é a falta de uma definição precisa de como essas explicações devem ser realizadas, algo que se deve especialmente ao fato de que pessoas diferentes em ambientes diferentes podem exigir diferentes tipos de explicações.

A inteligibilidade de sistemas de ML não deve consistir necessariamente em uma descrição precisa e detalhada de como os algoritmos funcionam, especialmente porque tal forma de fornecimento de informações pode ser inútil para o usuário final de uma plataforma de mídia social que recebe desinformação e que pretende aprender mais sobre como as informações são direcionadas para sua conta. Diferentemente, um especialista em auditoria de um sistema de reconhecimento facial provavelmente teria interesse em entender melhor seu código ou mesmo acessar os bancos de dados que o sistema utiliza para aprendizado, a fim de analisar se tal banco de dados é tendencioso ou não. Considerando que um nível ótimo de transparência depende da pessoa e do ambiente em que a decisão automatizada ocorre, é importante entender que os níveis e meios de fornecer informações variam de caso para caso.

Tal premissa representa um grande desafio para reguladores: como avaliar se o desenvolvedor ou implantador de um sistema de aprendizado de

máquina está fornecendo inteligibilidade suficiente e, portanto, cumprindo suas obrigações de transparência no processamento de dados pessoais e, assim, permitir que o titular dos dados exerça direitos sob a Lei Geral de Proteção de Dados (LGPD)?

Regulamentar um conjunto tão amplo de tecnologias compreende a necessidade de avaliar os sistemas aplicados a diferentes setores sociais e econômicos. Além disso, um mesmo sistema pode ser aplicado em diferentes mercados, o que significa que uma abordagem baseada no risco deve levar em consideração o tipo de aplicativo que está sendo implantado e a área em que é aplicado. Tal noção também leva à compreensão de que diferentes reguladores podem, portanto, ter parâmetros diferentes sobre o que seria informação suficiente sobre um determinado sistema de ML.

Com isso em mente, talvez o maior desafio seja determinar o que deve ou não ser explicado, uma vez que fornecer informações excessivas ou mesmo transmiti-las de forma inadequada tornaria a inteligibilidade do sistema ineficaz e desnecessariamente onerosa. Identificar como um sistema de ML deve ser compreensível requer, portanto, um olhar atento do regulador, que deverá avaliar diferentes aplicações e decidir que tipo de informação deve ser fornecida em um determinado caso concreto.

Para lançar luz sobre essa questão, este artigo tem como objetivo analisar a adequação da teoria da regulação responsiva na promoção da inteligibilidade do ML, investigando a adequação das suas estratégias participativas e o quadro de incentivos que proporciona para promover a explicabilidade.

INTELIGIBILIDADE DE SISTEMAS DE *MACHINE-LEARNING*

Desambiguações

A literatura sobre o fornecimento de informações sobre sistemas de aprendizado de máquina usa diferentes nomes para se referir à promoção de transparência. Devido às diversas definições utilizadas para descrever esses mecanismos, uma breve descrição de o que este artigo compreende sobre cada termo é fundamental para a realização desta discussão.

Este artigo usará os termos de inteligibilidade, transparência e compreensibilidade de forma intercambiável para se referir a quão compreensível é um sistema. Um sistema inteligível, transparente ou compreensível é aquele que permite ao ser humano compreender o seu funcionamento de forma a esclarecer dúvidas específicas sobre como é que executa determinadas decisões ou previsões. Nesse sentido, este trabalho aplicará esses conceitos sempre que não se referir a técnicas específicas, mas sim ao grau em que um sistema pode ser compreendido por um observador

(sistema mais inteligível, menos inteligível) ou se suas características podem ser compreendidas ou não (sistema inteligível, não inteligível).

Por outro lado, interpretabilidade e explicabilidade são conceitos que, embora amplamente utilizados por pesquisadoras, ainda não encontraram seu caminho para um consenso mais amplo. Alguns autores, como Tim Miller, equiparam os dois termos para se referir ao grau em que um observador pode compreender a causa de uma decisão e definem a explicação como um modo em que um observador pode compreender, mas, claramente, existem modos adicionais que se pode adotar, como tomar decisões que são inerentemente mais fáceis de entender ou via introspecção (MILLER, 2018, p. 14), ou seja, sistemas inerentemente transparentes e que não precisam de informações adicionais, vindas de sistemas adicionais.

No entanto, este artigo se baseia na posição de Cynthia Rudin (2018), que postula que existem diferenças entre explicabilidade e interpretabilidade. Ela argumenta que a explicabilidade seria uma forma de tornar as caixas pretas compreensíveis com ferramentas externas, o que ela chama de "explicações". A interpretabilidade, por sua vez, estaria relacionada a sistemas intrinsecamente compreensíveis, não demandando explicações adicionais sobre seus mecanismos. Essa perspectiva é mais adequada, especialmente quando se refere a técnicas específicas para promover inteligibilidade.

Uma outra nota conceitual deve ser feita em relação aos componentes dos sistemas de aprendizado de máquina. Como o conceito de algoritmo não abrange os dados aplicados para alimentar o sistema, mas apenas o conjunto de comandos que leva à sua saída, iremos nos referir ao termo sistema para englobar tanto os dados quanto o conjunto de instruções que o regem (algoritmo). Conforme discutiremos mais adiante, avaliar como os dados são processados por um sistema é especialmente importante para cumprir os regimes de proteção de dados, como o LGPD no Brasil e o RGPD na Europa.

Taxonomia de interpretabilidade e explicabilidade

Os métodos para promover a inteligibilidade podem ser classificados de acordo com diferentes critérios. Um deles refere-se ao momento em que o método é aplicável no que diz respeito à construção do modelo ML: antes (pré-modelo, *pre-model*), durante (no-modelo, *in-model*) ou depois (pós-modelo, *post-model*) do seu desenvolvimento.

As técnicas de explicação pré-modelo são desenvolvidas antes da construção do próprio modelo e, portanto, são independentes dele, aplicando-se apenas aos dados que serão usados para alimentar o sistema. Referem-se a técnicas de visualização de dados (CARVALHO e PEREIRA, 2019), como o t-

SNE, por exemplo, que permite a visualização de dados em alta dimensão ao dar a cada ponto de dados (data point) uma localização em um mapa bidimensional ou tridimensional (MAATEN e HINTON, 2008), ou a chamada Principle Component Analysis (JOLLIFFE, 2018). Do ponto de vista dos regimes de proteção de dados, tal abordagem é importante para avaliar quais dados estão sendo processados pelo sistema. No entanto, não permite, por si só, entender como tais dados estão sendo usados para alcançar decisões específicas no sistema.

As abordagens no-modelo, por outro lado, referem-se a modelos que incorporam ferramentas para explicar suas funcionalidades desde o seu próprio desenvolvimento, sendo, portanto, intrinsecamente interpretáveis. Visam responder à questão de como funciona o modelo (GILPIN, BAU, YAN, 2019) e, consequentemente, como processam os dados de treinamento.

As técnicas pós-modelo, por outro lado, dizem respeito à melhoria da explicabilidade de um sistema depois que ele já foi construído. (CARVALHO, DV; PEREIRA, EM; CARDOSO, 2019) A maioria das técnicas pós-modelo também são *post-hoc*, o que significa que o modelo é explicado depois de já ter sido treinado e visa responder à pergunta o que mais o modelo pode dizer. Segundo Lipton, uma vantagem desse conceito de interpretabilidade é que podemos interpretar modelos opacos após o fato, sem sacrificar o desempenho preditivo (LIPTON, 2016).

As técnicas de promoção de inteligibilidade também podem ser categorizadas de acordo com seu escopo, que refere-se à parte do processo de predição que pretendem explicar (CARVALHO.; PEREIRA; CARDOSO, 2019). Elas podem fornecer transparência algorítmica ou interpretabilidade global e local.

A transparência algorítmica permite compreender como o algoritmo aprende com os dados e que tipo de relações pode extrair de tal operação. Nesse sentido, o objetivo da transparência algorítmica é aprender como o algoritmo funciona, e não previsões individuais. Não requer conhecimento sobre os dados ou o modelo aprendido, mas estritamente sobre o próprio algoritmo, ou seja, o conjunto de instruções que permite ao sistema realizar uma tarefa específica. Portanto, é uma forma de responder à pergunta "como o modelo treinado faz previsões?" (MOLNAR, 2019).

Por sua vez, a interpretabilidade global é aplicada quando o objetivo do agente é descrever o comportamento de todo o modelo, o que inclui uma compreensão dos dados e do próprio algoritmo (ARYA et al, 2019). Esta explicação pode ser holística ou modular. A interpretabilidade do modelo holístico global visa explicar todo o modelo de uma vez e entender como ele faz previsões, o que requer conhecimento do algoritmo e dos dados de treinamento.

Um modelo só pode ser holístico se for simples o suficiente, uma vez que qualquer modelo que tenha mais de 5 parâmetros ou pesos provavelmente não caberá na memória de curto prazo do ser humano médio (COWAN, 2010) e, portanto, é muito difícil de ser alcançado (MOLNAR, 2019).

Já a interpretabilidade do modelo global no nível modular é mais praticável de ser alcançada. Ela não tem como objetivo explicar todos os recursos de um modelo de aprendizado de máquina, mas, em vez disso, explicar o modelo separando recursos específicos usados nos processos de tomada de decisão e tentando entender como eles funcionam. A pergunta a ser respondida por este método é "como as partes do modelo afetam as previsões?" (MOLNAR, 2019)

Finalmente, a interpretabilidade local visa descrever previsões únicas e pode ser alcançada explicando (1) uma única previsão, examinando o que o modelo previu após o processamento de um dado de entrada específico e explicar por que; ou explicando (2) um grupo de previsões, selecionando um grupo de instâncias e entendendo como o modelo faz previsões específicas para este grupo (ARYA et al, 2019).

A utilidade de explicações globais ou locais depende das informações de que um indivíduo precisa para atingir um objetivo específico compreendendo o sistema que deseja. Por exemplo, para entender qual papel um sistema de personalização de conteúdo desempenha em uma rede social para traçar o perfil de usuários e exibir conteúdo personalizado, um regulador provavelmente faria melhor uso das explicações do modelo global. O objetivo do regulador seria principalmente compreender como funciona o sistema em geral, de forma a desenvolver uma regulação mais eficaz para orientar plataformas no desenvolvimento de algoritmos que melhor identifiquem a desinformação online e respondam rapidamente, por exemplo, reduzindo o alcance de um conteúdo específico. A transparência do algoritmo também pode ser útil neste contexto, uma vez que o entendimento da cadeia de comandos pode se mostrar útil para identificar eventuais vieses em suas métricas (BOZDAG, 2013).

Por outro lado, um usuário que não deseja receber publicidade específica em um mecanismo de busca provavelmente faria melhor uso de um sistema que explica como direcionou esse conteúdo para o usuário, com base em quais entradas e o peso de cada uma delas em essa recomendação particular. Nesse sentido, uma explicação local possivelmente seria mais adequada.

Uma explicação local também pode ser útil para avaliar se um sistema de pontuação de crédito foi discriminatório ou não ao avaliar uma pessoa específica. Entender quais dados foram usados e como o sistema se saiu para tomar essa decisão específica podem ser aspectos para os reguladores e os usuários examinarem.

Outra taxonomia para explicação está relacionada ao grau em que um usuário pode interagir com um sistema, que pode ser estático ou interativo. Uma explicação estática não muda em resposta ao *feedback* do usuário. Por outro lado, uma interativa permite que os usuários dialoguem com o modelo para solicitar mais informações sobre uma decisão tomada, como detalhando ou pedindo diferentes tipos de explicações (por exemplo, por meio do diálogo) até que fiquem satisfeitos (ARYA et al , 2019).

Tendo observado como as explicações geralmente são fornecidas, passamos à questão de por que as explicações podem ser uma ferramenta útil para fornecer aos indivíduos mais controle sobre como os sistemas estão afetando suas vidas com base no processamento de dados pessoais. Assim, recorremos aos regimes de proteção de dados no Brasil e na Europa para avaliar como eles regulamentam esse tema.

EXPLICABILIDADE DE SISTEMAS DE *MACHINE-LEARNING* E REGIMES DE PROTEÇÃO DE DADOS - UM DIREITO À EXPLICAÇÃO?

Tornar sistemas transparentes e compreensíveis não é apenas uma questão de informar um indivíduo sobre o funcionamento de um sistema por si só. Também pode ser visto como uma forma de proteger dados pessoais e abordar a discriminação algorítmica. Esta sessão irá, portanto, explorar como a transparência dos sistemas de decisão automatizada (termo utilizado tanto pelo regime europeu quanto pelo brasileiro e que inclui sistemas de *machine-learning*) é abordada tanto na legislação europeia quanto na brasileira de proteção de dados, a fim de avaliar se elas fornecem um direito a explicação.

O principal objetivo dos regimes de proteção de dados é fornecer aos usuários controle sobre seus dados. Essa é a ideia central que permeia o direito à autodeterminação informacional, tal como estabelecido pela Corte Constitucional Alemã em 1983 (MAYER-SCHÖNBERGER, 1997). Uma vez que nos concentramos aqui em sistemas cujas decisões envolvem o processamento de dados pessoais, tais atividades são frequentemente abrangidas por leis de proteção de dados. Analisaremos tanto a Lei Geral de Proteção de Dados brasileira (LGPD) quanto o Regulamento Geral de Proteção de Dados europeu (RGPD), a fim de desenvolver uma perspectiva comparativa de como regimes semelhantes lidam com o tema da transparência dos sistemas automatizados de tomada de decisão. Começamos pelo RGPD, já que a União Européia tem um *corpus* jurídico de análise mais substancial sobre eles do que o Brasil, que tem uma experiência muito mais recente em proteção de dados.

O Regulamento Geral Europeu de Proteção de Dados (RGPD) estabelece disposições específicas relacionadas à transparência dos sistemas automatizados

de tomada de decisão. Para isso, cria o direito de o titular dos dados receber informações específicas sobre esse tipo de tomada de decisão e o sistema responsável por ela.

Nos termos dos artigos 13(2)(f) e 14(2)(g), o RGPD estabelece que as seguintes informações devem ser fornecidas espontaneamente pelo controlador de dados ao titular dos dados, caso haja um processamento de dados pessoais por uma decisão automatizada:

1. A existência de tomada de decisão automatizada;
2. Informações significativas sobre a lógica do sistema;
3. O que significa e as consequências previstas de tal tratamento para o titular de dados.

De acordo com o RGPD, todas essas informações devem ser fornecidas de forma proativa pelo controlador de dados, sem a necessidade de solicitação do titular dos dados. Por essa abordagem, o RGPD visa proteger os titulares dos dados especialmente no que diz respeito às decisões automatizadas que são realizadas para fins de criação de perfis (perfilização, *profiling*) com base no processamento de dados sensíveis, algo que expressaria um grande risco para a proteção de dados.

O RGPD é rígido quanto às decisões automatizadas que devem ser permitidas ao lidar com o processamento de dados pessoais. Nos termos do seu considerando 71, descreve que a tomada de decisões automatizada que avalie, por exemplo, os aspectos pessoais relativos a uma pessoa singular quando produz efeitos jurídicos só deve ser efetuada quando expressamente autorizada pelo direito da União Europeia ou dos Estados-Membros. Esse processamento deve estar sempre sujeito a salvaguardas adequadas, que devem incluir informações específicas do titular dos dados e o direito de obter intervenção humana, de expressar o seu ponto de vista, de obter uma explicação da decisão tomada após tal avaliação e de contestar a decisão.

Embora o RGPD destaque a sensibilidade das decisões automatizadas sobre um indivíduo e como elas podem impactar o exercício dos direitos por um indivíduo, ele não é muito específico sobre como uma explicação deve ser feita e quais informações devem ser fornecidas pelo sistema. Também há desacordo sobre se o regulamento exige que os controladores expliquem apenas a lógica subjacente a esses sistemas ou também decisões específicas e, nesses casos, quais decisões devem ser explicadas.

Por exemplo, Wachter, Mittelstadt e Floridi (2017) argumentam que o RGPD estabelece não o direito à explicação das decisões individuais, mas sim um mero "direito a ser informado". Esse direito englobaria informações relativas apenas à lógica geral envolvida no sistema, bem como a importância e as consequências previstas de sistemas automatizados de tomada de decisão. Se

aplicarmos a taxonomia fornecida na seção anterior em relação aos diferentes métodos de explicabilidade do sistema, tal interpretação levaria ao entendimento de que o RGPD obriga os controladores a fornecer apenas interpretabilidade do modelo a nível global.

Na direção oposta, Selbst e Powles defendem que o RGPD deve ser lido como estabelecendo um direito à explicabilidade, na medida em que o titular dos dados deve receber informações suficientes para exercer seus direitos. Em outras palavras, toda vez que os direitos de um titular de dados forem colocados em risco devido ao processamento de dados pessoais por um sistema automatizado de tomada de decisão, qualquer explicação necessária para avaliar se há de fato uma violação do RGPD deve ser fornecida. Isso incluiria, por exemplo, meios para identificar se uma decisão automatizada específica foi tendenciosa ou não. Portanto, não haveria qualquer restrição *a priori* sobre o método de explicabilidade a ser aplicado, desde que fornecesse informações suficientes para o exercício de seus direitos de proteção de dados (SELBST e POWLES, 2017).

A interpretação de Selbst e Powles para o RGPD parece bastante adequada, pois engloba uma abordagem sistemática para a regulamentação. Ao abordar o processamento de dados por sistemas automatizados de tomada de decisão, o Considerando 38 do RGPD afirma que a perfilização que resulta em discriminação contra pessoas físicas com base em dados pessoais que são, por sua natureza, particularmente sensíveis em relação aos direitos e liberdades fundamentais, deve ser proibida nas condições estabelecidas nos Artigos 21 e 52 da Carta. O artigo 11 do RGPD complementa esse raciocínio ao postular que a perfilização que resulte na discriminação de pessoas singulares com base em categorias especiais de dados pessoais a que se refere o artigo 10º é proibida, nos termos do direito da União.

Seria difícil avaliar se uma decisão automatizada específica com base em atividades de criação de perfil pode ter sido tendenciosa sem analisar a decisão (*output*) específica do sistema, e não apenas a lógica subjacente do modelo de aprendizado de máquina como um todo. Nesse sentido, parece mais apropriado argumentar que o RGPD aborda o direito à explicabilidade, pelo menos em casos de definição de perfil, e especialmente aqueles realizados por meio do processamento de dados sensíveis.

A Lei Geral de Proteção de Dados (LGPD) brasileira, por outro lado, também trata do direito de solicitar informações sobre o funcionamento de sistemas automatizados de tomada de decisão, incluindo aqueles destinados a definir o perfil pessoal, profissional, de consumo ou de crédito ou aspectos da personalidade do indivíduo, nos termos do seu Artigo 20. Diferentemente do RGPD, não há obrigação específica da LGPD para que o controlador forneça

tais informações de forma proativa, mas apenas mediante solicitação do titular dos dados. Além disso, embora exista a obrigação de os titulares dos dados solicitarem revisões de decisões tomadas por sistemas automatizados, não há nenhum requisito específico de que essa revisão seja feita por um agente humano.

De acordo com o artigo 20, §1º, da LGPD, "§ 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". Caso o responsável pelo tratamento não forneça informações argumentando que tal infringiria o sigilo comercial e industrial, a LGPD estabelece que a autoridade nacional pode efetuar uma auditoria para verificar aspectos discriminatórios específicos no tratamento automatizado de dados pessoais (art. 20, §2º, LGPD).

Neste sentido, a LGPD deixa espaço suficiente para que os sistemas sejam auditados e para que os controladores sejam obrigados a prestar esclarecimentos a pedido de uma pessoa em causa. Como é o caso do RGPD, embora não haja certeza de que os controladores terão que divulgar explicações sobre decisões específicas tomadas por seus sistemas automatizados, pelo menos explicações do modelo global provavelmente deverão ser fornecidas, uma vez que a transparência no processamento de dados é um princípio em ambas as legislações (Artigo 6, VI, LGPD e Artigo 5 (1) (a), RGPD).

Em todo caso, levando em consideração que a redação da LGPD, mais uma vez semelhante ao RGPD, apresenta um grau considerável de incerteza sobre como e para que fins as informações sobre os "critérios e procedimentos utilizados para uma decisão automatizada" devem ser fornecidas, cabe tanto a controladores de dados quanto a reguladores pressionar por mais diálogo, a fim de avaliar como abordar a transparência de sistemas de ML de uma forma que forneça mais controle ao titular dos dados sem afetar substancialmente a usabilidade desses sistemas.

A quantidade de informações que terá de ser fornecida pelo controlador de dados sobre o processamento de dados terá, portanto, de ser avaliada caso a caso. Considerando que tanto a LGPD como o RGPD fazem referência a meios específicos de combate às atividades de criação de perfis que possam conduzir a práticas discriminatórias, a autoridade deve fazer, assim, uma análise do risco envolvido nesse tratamento de dados. Para mergulhar mais fundo em tal análise de risco, a próxima seção se debruça sobre a abordagem baseada em risco, adotada por ambas as legislações.

LGPD, RGPD E A ABORDAGEM BASEADA EM RISCO

Tanto o LGPD quanto o RGPD são frequentemente mencionados como tendo introduzido uma abordagem baseada em risco para fazer cumprir suas obrigações. Isso está expresso na LGPD, por exemplo, em seu artigo 44, II, que estabelece que

> 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.

No RGPD, por sua vez, uma abordagem semelhante é encontrada em suas obrigações de privacidade desde a concepção e por padrão (*privacy-by-design* e *by-default*), que estimula os controladores de dados a adotar as salvaguardas necessárias para o processamento de dados de acordo com, entre outras variáveis, os riscos envolvidos na operação (artigo 25.º, n.º 1, RGPD).

Com foco no risco, os regimes de proteção de dados permitem a construção de um espaço de confiança mútua entre os controladores e a autoridade, em que os controladores recebem um voto de confiança para definir quais das atividades que realizam representam um risco maior, sendo assim obrigados fornecer meios mais substanciais para proteger os dados pessoais.

O grau de risco e as normas de proteção e segurança de dados que são postas em prática pelo controlador para o processamento de dados, o que inclui a adoção de medidas de privacidade desde a concepção, também são de grande importância, uma vez que são um dos aspectos a serem avaliados por autoridades em caso de violação de dados ou qualquer outra violação dos direitos de proteção de dados.

Nesse sentido, de acordo com a LGPD, artigo 48, §3º, "[n]o juízo de gravidade do incidente, será avaliada eventual comprovação de que foram adotadas medidas técnicas adequadas que tornem os dados pessoais afetados ininteligíveis, no âmbito e nos limites técnicos de seus serviços, para terceiros não autorizados a acessá-los." Da mesma forma, o RGPD estabelece que, ao avaliar a imposição de multas por infrações, as medidas de segurança e privacidade desde a concepção devem ser levadas em consideração pela autoridade de supervisão (Artigo 83 (2) (d), RGPD).

É, portanto, apropriado concluir que ambas as normas de proteção de dados preveem diferentes níveis de exigências em função dos riscos envolvidos nos tratamentos de dados. Em relação aos sistemas automatizados de tomada de decisão, ou, mais especificamente, modelos de aprendizado de máquina, a explicabilidade - ou transparência como um todo - deve ser vista como uma das

medidas a serem implementadas pelo controlador de forma a enfrentar os riscos do processamento de dados, especialmente quando o sistema é responsável por tomar decisões de alto risco. Afinal, como afirma Gonçalves, com o advento do *big data*, muitas vezes não é a coleta de informações em si que é sensível, mas as inferências inerentemente obscuras que são extraídas dela e a maneira como essas inferências são extraídas (GONÇALVES, 2019).

É razoável considerar, pelo menos no âmbito do RGPD, que as atividades de criação de perfil realizadas por sistemas automatizados são arriscadas, pois o regulamento é específico que um Relatório de Impacto à Proteção de Dados (RIPD) deve ser fornecido pelo controlador em relação a decisões "que produzam efeitos jurídicos em relação à pessoa física ou similar afetam significativamente a pessoa singular "e que processa em larga escala categorias especiais de dados (Artigo 35 (3)).

A LGPD silencia sobre o assunto, mas há espaço para a Autoridade Nacional de Proteção de Dados (ANPD) tratar das situações em que o RIPD seja obrigatório. Considerando que o citado artigo 20, LGPD, trata especificamente de atividades de definição de perfis que possam levar à discriminação, não seria surpresa se a ANPD exigisse RIPDs para sistemas responsáveis por tais formas de processamento de dados. Nesse sentido, fornecer explicações seria uma ferramenta útil para o cumprimento por parte do controlador de dados e para a aplicação da legislação pela autoridade de supervisão. Isso será mais aprofundado adiante, quando falaremos sobre a aplicação de RIPDs sobre sistemas de inteligência artificial.

Com isso em mente, cabe agora abordar o que deve ser explicado com relação a esses modelos.

EXPLICAR O QUÊ?

Independentemente de o RGPD ou a LGPD estabelecerem explicitamente o direito à explicação, a crescente dependência de sistemas de aprendizado de máquina para a realização de decisões de alto risco em áreas como saúde e aplicação da lei, bem como o efeito crescente das campanhas de desinformação em várias democracias, pode inevitavelmente exigir mais demandas de transparência algorítmica. No Brasil, por exemplo, as demandas por mais transparência podem ocorrer especialmente à medida que a implantação de tecnologias de reconhecimento facial aumenta exponencialmente. Além disso, como os debates a respeito da chamada "Lei das *Fake News*" (PL n. 2.630/2020) têm mostrado que o combate à desinformação está no topo da agenda regulatória, a transparência dos sistemas algorítmicos de sistemas de personalização de conteúdo já está em debate.

Do ponto de vista regulatório, não seria proporcional exigir uma explicação para cada decisão tomada por um modelo de ML. Explicações não são gratuitas. Conforme observado por Doshi-Velez et al, gerá-las leva tempo e esforço, reduzindo assim o tempo e o esforço disponíveis para gastar em outra conduta potencialmente mais benéfica. Portanto, a eficácia das explicações deve ser balanceada com o custo de gerá-las (DOSHI-VELEZ et al., 2017). Por esse motivo, qualquer regulamentação sobre o tema enfrenta o desafio de encontrar o equilíbrio entre o que é uma informação significativa a ser fornecida ao usuário e o que não é (RUDIN, 2019).

Conforme mencionado acima, tanto LGPD quanto RGPD estão especialmente preocupados com decisões automatizadas que podem levar à discriminação do titular dos dados. Nesse sentido, um primeiro filtro para determinar quais decisões específicas devem ser explicadas incluiria aquelas atividades de criação de perfil realizadas por meios automatizados com base no processamento de categorias especiais de dados. Tais decisões automatizadas devem, a priori, estar sujeitas a explicações locais em ambos os sistemas jurídicos, caso possam levar a discriminação que afete o interesse do usuário.

Uma explicação deve permitir que o usuário compreenda como uma determinada entrada influenciou uma saída que teve efeitos sobre os interesses do titular dos dados, nos termos do Artigo 20, LGPD, ou que teve efeitos jurídicos sobre a pessoa em questão, nos termos do Artigo 22 (1), RGPD .

Uma explicação útil deve ser capaz de fornecer as informações a seguir (DOSHI-VELEZ et al., 2017):

1. Quais foram os principais insumos levados em consideração em uma decisão? Idealmente, uma resposta os apresentaria em ordem de significância, a fim de entender se dados sensíveis específicos foram levados em consideração para uma dada decisão;
2. Mudar um determinado fator mudaria a decisão? Isso permitirá ao usuário entender se um fator específico foi determinante para a predição feita pelo modelo de *machine-learning*;
3. O sistema pode tomar decisões diferentes em dois casos semelhantes? Isso permite que um ser humano entenda se a mesma variável teria pesos diferentes em casos semelhantes.

Essas questões são essenciais para garantir que saibamos quais fatores foram decisivos em um processo automatizado de tomada de decisão e, consequentemente, para identificar os vieses de um sistema algorítmico de personalização de conteúdo. Eles permitem um entendimento completo de por que certas informações foram mostradas a um usuário e quais informações não foram mostradas.

Por outro lado, Doshi-Velez et al. (2017) também fornecem três outros componentes para identificar quais decisões requerem explicação:

1. A decisão deve ter um impacto sobre uma pessoa que não seja o tomador de decisão;
2. Deve haver valor em saber se a decisão foi tomada erroneamente, como para responsabilizar o tomador de decisão pelos danos que a decisão possa ter causado;
3. Deve haver algum motivo para acreditar que ocorreu (ou ocorrerá) um erro no processo de tomada de decisão.

Bayamlioglu (2018) acrescenta que a precisão, a adequação das entradas com as saídas, bem como a metodologia aplicada para as previsões devem ser divulgadas por modelos de ML. Além disso, a informação sobre o contexto em que a decisão foi tomada também é crucial - diz respeito a fornecer informações ao usuário sobre onde a decisão começa e termina, permitindo uma avaliação do modelo juntamente com o impacto posterior da decisão em vista das finalidades declaradas e não declaradas do sistema.

Considerando que geralmente há falta de conhecimento técnico entre formuladores de políticas e reguladores sobre temas como análise de dados ou mesmo sistemas de aprendizado de máquina como um todo, é de extrema importância criar ferramentas para avaliar se uma técnica de explicação é útil ou não. Para lançar uma luz sobre esta questão, devemos primeiro nos mover em direção a um entendimento sobre quais são os objetivos da explicabilidade.

Rüping (2006) argumenta que um sistema explicável deve perseguir três objetivos principais. O primeiro é a precisão, que se refere a ter uma conexão entre a previsão feita pelo modelo de aprendizado de máquina e a explicação fornecida pelo método de explicação. É importante porque é possível ter uma hipótese compreensível que não tenha conexão com os dados. A segunda é a compreensibilidade, referindo-se à facilidade de compreensão de uma explicação pelo usuário. Principalmente no que diz respeito à explicabilidade dos sistemas de recomendação, como em uma rede social ou em um mecanismo de busca, por exemplo, a maioria dos usuários não é hábil em computação, a explicação de uma decisão deve ser fácil de entender para ser minimamente útil. Em terceiro lugar, um método explicável deve ser eficiente, o que significa que reflete o tempo necessário para um usuário compreender totalmente a explicação. Refere-se, assim, a quão compreensível é a explicação em um período finito e preferencialmente curto (CARVALHO, PEREIRA, CARDOSO, 2019).

Há pouca pesquisa sobre como avaliar os métodos de explicação, tornando difícil para os reguladores identificarem como eles podem medir se uma técnica específica é mais adequada para uma determinada aplicação, ou se

permite uma compreensão suficiente da lógica do sistema ou de decisão específica (CARVALHO, PEREIRA, CARDOSO, 2019). Considerando que nem todos os modelos são igualmente explicáveis e que nem todas as aplicações têm as mesmas necessidades de explicabilidade, é importante desenvolver métodos baseados em evidências em métodos de mensuração de explicabilidade (DOSHI-VELEZ e KIM, 2017). Doshi-Velez e Kim propõem uma série de questões a serem realizadas a fim de identificar quais as principais características que um sistema explicativo deve ter em relação ao contexto que se pretende explicar (DOSHI-VELEZ e KIM, 2018).

A primeira diz respeito a um ponto já mencionado acima: é preciso entender todo o sistema ou uma decisão específica? No primeiro caso, a interpretabilidade global seria necessária, enquanto no último as explicações locais pareceriam suficientes. Com relação aos sistemas de recomendação, por exemplo, objetivos diferentes tornariam ambas as abordagens úteis: se alguém precisa entender como um algoritmo classifica o conteúdo de acordo com o envolvimento dos usuários com eles de uma forma mais generalizada, uma explicação global seria necessária. No entanto, se um usuário pretende saber por que ele teve acesso a um dado específico de desinformação sobre uma cura milagrosa para COVID-19, ele provavelmente acharia mais interessante saber quais dados pessoais o sistema tomou como parâmetro (*input*) para chegar à exibição do conteúdo desinformativo (*output*).

A segunda questão seria a caracterização da incompletude, que visa responder que parte da formulação está incompleta e em que medida. Isso se relaciona a qual explicação é necessária, como saber sobre o conjunto de dados usado em um aplicativo de reconhecimento facial que parece ter feito uma identificação tendenciosa, ou caso seja necessário saber sobre a imagem capturada pela câmera do sistema. Cada caso exigirá uma forma diferente de explicação.

As restrições de tempo: quanto tempo o usuário pode gastar tentando entender a explicação? Para saber por que o conteúdo desinformativo foi exibido para ela, o usuário pode querer gastar no máximo dois minutos para entender a explicação. Um pesquisador em busca de conjuntos de dados tendenciosos em aplicativos de reconhecimento facial provavelmente estaria interessado em passar horas analisando dados.

Por fim, o grau e a natureza de conhecimento do usuário: qual é a experiência do usuário para entender esse tipo de explicação? Para um médico que tenta entender por que um modelo de ML classificou um tumor com base na imagem de um paciente, provavelmente se exigiria um nível de sofisticação do método de explicação diferente do usuário da rede social que não precisa de um entendimento complexo da natureza do perfil sendo realizado sobre ela. O

tipo de linguagem exigida também seria diferente nesses exemplos: enquanto o médico acharia crucial receber as informações em um texto técnico, uma linguagem mais simples certamente seria mais adequada para o usuário da rede social.

Os quatro aspectos aqui destacados podem se mostrar úteis para os reguladores na hora de analisar se e como uma determinada informação deve ser fornecida para aumentar a inteligibilidade de um sistema de aprendizado de máquina. Portanto, lançamos mão agora de uma análise de como a teoria regulatória pode apoiar reguladores a promoverem maior transparência em sistemas de ML.

UMA TAREFA PARA A REGULAÇÃO RESPONSIVA?

Panorama teórico da regulação responsiva

Como discutimos anteriormente, a opacidade dos sistemas de *machine-learning* (ML) representa grandes ameaças ao exercício de direitos e liberdades. Neste sentido, sob uma perspectiva de proteção de dados, a avaliação de qual grau de transparência de um sistema de ML é relevante para permitir o exercício dos direitos dos titulares dos dados varia de caso para caso. Algumas características que devem ser analisadas pelos reguladores abrangem, por exemplo, os riscos que um sistema representa quanto ao setor em que será aplicado ou à categoria de dados pessoais processados. Quanto maior o risco de um sistema, provavelmente maior também será o grau de transparência necessário para evitar os danos potenciais em que o sistema pode incorrer.

Muito debate tem sido feito sobre se os sistemas de inteligência artificial devem ou não ser regulamentados (BLACK e MURRAY, 2019). No entanto, considerando os riscos envolvidos nesses sistemas, a questão não deve se estender no debate simplista de regulamentar-desregulamentar, mas sim na questão de como a regulamentação de IA deve ser projetada, especialmente no que diz respeito à inteligibilidade.

Além disso, dado o caráter rapidamente transformador da tecnologia de ML, ter uma compreensão completa da dinâmica dos mercados regulamentados é crucial para os reguladores impulsionar estratégias regulatórias eficazes, especialmente quando aplicadas a sistemas de ML (BLACK e MURRAY, 2019, p. 12).

Como resultado, uma regulação eficaz deve ser capaz de, por um lado, permitir a maleabilidade do regulador para implementar diferentes políticas para abordar diferentes sistemas de ML em diferentes contextos e, por outro lado, um diálogo constante não apenas entre regulador e regulados, mas também com outras partes interessadas.

A teoria da regulação responsiva veio à luz com o objetivo exato de transcender o impasse entre aqueles que defendem mais regulamentação e aqueles que são a favor da desregulamentação. Ao argumentar que uma boa política regulatória trata de compreender a regulamentação privada - por associações da indústria, por empresas, por pares e por consciências individuais - e como ela é interdependente com a regulamentação estatal, Ayres e Braithwaite propuseram, em seu livro *Responsive Regulation: Transcending the Deregulation Debate*, que, na maioria dos casos, a mistura entre a regulação pública e privada abriu possibilidades efetivas de abordagem das questões socioeconômicas que surgem em diferentes mercados (AYRES e BRAITHWAITE, 1992, p. 3).

Muito trabalho foi feito por outras pesquisadoras em torno do primeiro livro de Ayres e Braithwaite, mas suas premissas principais persistem. Gostaríamos de destacar duas delas. A primeira diz respeito a uma cooperação necessária entre reguladores e regulados, uma característica crucial na qual a teoria se baseia para promover a conformidade por meio de negociação efetiva entre empresas e agências estatais. Esse diálogo leva a um melhor entendimento do mercado e também a uma maior confiança entre esses atores, o que aumenta sua capacidade de cooperação. Mas por que a cooperação é importante?

A regulação responsiva trata principalmente de encontrar o equilíbrio certo entre punição e persuasão. Segundo sua lógica, por um lado, quando reguladores adotam estratégias baseadas somente na punição, suas ações prejudicam a boa vontade de atores que sejam motivados por um senso de responsabilidade. Por outro lado, quando a estratégia é totalmente baseada na persuasão e na autorregulação, a ação do Estado provavelmente será explorada quando os atores forem motivados exclusivamente pela racionalidade econômica (BRAITHWAITE, 1985).

Os teóricos consideram os atores corporativos como feixes de compromissos contraditórios com valores sobre racionalidade econômica, cumprimento da lei e responsabilidade empresarial (AYRES e BRAITHWAITE, 1992, p. 19). Por isso, cada empresa exigirá abordagens diferentes do Estado para garantir o cumprimento, pois suas motivações diferem de uma para outra. Alguns agentes regulados estarão naturalmente mais dispostos a cumprir a lei e, portanto, abordagens persuasivas por meio de negociação podem ser mais adequadas para eles do que a imposição de sanções severas, que muitas vezes ignoram o histórico de compliance de seus atores e punem agentes bem-intencionados por uma violação legal ocasional. Para outros atores corporativos que consideram a lei um mero obstáculo para ganhos econômicos e estão constantemente tentando evitá-la, talvez a persuasão não seja eficaz e, portanto, requeira punições mais severas. Nesse sentido, a adoção

de uma estratégia *tit-for-tat* (TFT) que é tanto provocativa quanto indulgente tem mais probabilidade de ser eficaz (AYRES e BRAITHWAITE, 1992, p. 5).

Com isso em mente, os autores da teoria afirmam que não existem soluções regulatórias ótimas ou melhores, pois as abordagens eficazes podem variar de acordo com o mercado específico, o contexto histórico e os negócios envolvidos (AYRES e BRAITHWAITE, 1992, p. 5). O papel do regulador é, portanto, estar atento às diferentes características dos negócios e dos mercados, de modo a identificar quando e como agir, e como elaborar normas que sejam mais adequadas a essas diferentes realidades. Nesse sentido, os objetivos regulatórios são, segundo os autores, mais facilmente alcançados quando as agências apresentam tanto uma hierarquia de sanções quanto uma hierarquia de estratégias regulatórias de vários graus de intervencionismo (AYRES e BRAITHWAITE, 1992, p. 6). A intervenção do Estado nos negócios aumenta e diminui de acordo com o nível de cumprimento das entidades reguladas.

Isso nos leva à segunda premissa que gostaríamos de destacar: as pirâmides regulatórias. Ayres e Braithwaite usam figuras de pirâmides para ilustrar como a ação regulatória deve ser direcionada aos regulados tanto em termos de estratégias regulatórias quanto de imposição de sanções. Quanto mais correto do ponto de vista regulatório for um ator corporativo, menor será o grau de intervenção do Estado em suas atividades, habitando assim os níveis mais baixos das pirâmides. Porém, nos casos em que uma entidade passa a violar normas legais de forma recorrente, o regulador terá legitimidade para agir, impondo sanções mais severas e escalando o nível de intervenção, subindo na pirâmide (AYRES e BRAITHWAITE, 1992, p. 6). Os autores apresentam dois exemplos de pirâmides regulatórias, respectivamente uma de uma pirâmide de constrangimentos (AYRES e BRAITHWAITE, 1992, p. 35), e outra de estratégias regulatórias (AYRES e BRAITHWAITE, 1992, p. 39).

Para os teóricos, quanto maior o nível de fiscalização que os reguladores puderem escalar na pirâmide, maior será a disposição dos regulados em obedecer (AYRES e BRAITHWAITE, 1992, p. 6). Essa abordagem é a essência da estratégia *tit-for-tat* mencionada acima, em que o regulador se abstém de uma resposta dissuasiva enquanto a empresa estiver cooperando; mas quando a empresa cede à tentação de explorar a postura cooperativa do regulador e trapaceia no cumprimento, então o regulador muda de uma resposta cooperativa para uma resposta dissuasora (AYRES e BRAITHWAITE, 1992, p. 21). No entanto, é fundamental entender que, como as motivações e dinâmicas das empresas e dos mercados variam entre si, diferentes formas de *enforcement* e estratégias terão que ser concebidos pelo regulador, quase de uma maneira sob medida para cada ator corporativo.

Nos anos que se seguiram à publicação do livro de Ayres e Braithwate, outros recursos foram adicionados ao desenvolvimento da teoria. Dentre eles, podemos destacar (i) o diamante regulatório proposto por Kolieb por incluir na fundamentação da teoria mecanismos de recompensa aos regulados pela adoção de medidas que vão além do mero cumprimento da lei (KOLIEB, 2015); e (ii) a ideia de governança em rede, que se relaciona à criação de uma "sociedade reguladora" onde ONGs, órgãos de auditoria e pressão social local teriam um papel fundamental nos esforços regulatórios, especialmente em países em desenvolvimento (BRAITHWAITE, 2016). Tal quadro participativo permite que diversos grupos de interesse participem e forneçam insumos na dinâmica regulatória, garantindo que os reguladores e regulados atendam de forma mais adequada às necessidades dos diferentes grupos sociais.

Regulação responsiva aplicada à inteligibilidade de sistemas de aprendizagem de máquina

Tendo apresentado os principais fundamentos da teoria da regulação responsiva, agora é hora de questionar se e como ela deve ser aplicada aos sistemas de ML e, mais especificamente, para regular a sua inteligibilidade.

Avaliamos anteriormente como a Lei Geral de Proteção de Dados (LGPD) se aplica à explicabilidade em sistemas automatizados de tomada de decisão, que abrangem aprendizado de máquina. Nesse sentido, avaliar como a lei permitiria ao seu principal regulador, a Autoridade Nacional de Proteção de Dados (ANPD), aplicar a teoria da regulação responsiva seria um primeiro passo adequado.

Segundo Renata Garcia (GARCIA, 2020), duas características principais da LGPD podem ser avaliadas como uma porta de entrada para a aplicação da regulação responsiva. O primeiro diz respeito ao caráter participativo da legislação. Por exemplo, nos termos do artigo 55-J, §2º, a lei obriga a ANPD a ouvir demandas de diferentes grupos de interesse no desenvolvimento de nova regulamentação. Além disso, uma das principais disposições para abordar a cooperação entre o regulador e os regulados é aquela que descreve o encarregado de proteção de dados (DPO), que é a "pessoa indicada pelo controlador e operador para atuar como canal de comunicação entre o controlador, os titulares dos dados e a Autoridade Nacional de Proteção de Dados (ANPD)" (art. 5º, VIII, LGPD). O DPO é crucial para promover a criação de um ambiente confiável para aumentar a cooperação sob uma lógica de regulação responsiva, monitorando a conformidade de sua organização com a LGPD e emitindo recomendações tanto à empresa quanto a seus funcionários (IRAMINA, 2020, p. 107).

Ainda nesse sentido, outra característica que permite maior participação de setores interessados alheios à dualidade direta regulador-regulado para a aplicação da LGPD é o Conselho Nacional de Proteção de Dados Pessoais e da Privacidade (CNPD). Descrito pelo art. 58-A e 58-B da lei, esse corpo consultivo, ainda em vias de formação quando da realização deste trabalho, será composto por 23 integrantes de setores como os poderes Executivo e Legislativo federais, academia, sociedade civil, confederações sindicais, entidades representativas do setor empresarial, Comitê Gestor da Internet, dentre outros. Suas atribuições incluem o papel de aconselhar a ANPD em possíveis ações a serem tomadas por ela, incluindo pela propositura de diretrizes estratégicas, bem como de elaborar estudos e disseminar o conhecimento sobre a proteção de dados pessoais e da privacidade à população. Tudo isso pode representar mais um meio para interpretar a LGPD como abarcante a noção de governança em rede proposta por Braithwaite (2016).

A segunda característica que aproxima a LGPD de uma regulamentação responsiva é o seu artigo 52, que prevê uma pirâmide de fiscalização, em que penalidades que vão desde advertências e multas até a proibição de atividades de processamento de dados podem ser aplicadas de "forma gradativa, isolada ou cumulativa". A escalada da intervenção regulatória dependerá, portanto, do comportamento dos regulados e dos resultados obtidos (GARCIA, 2020, p. 55).

Esses dois aspectos da LGPD são elementos que, de antemão, abrem espaço para a aplicação da teoria responsiva. Ainda, deve-se destacar que a LGPD, em seu art. 50, incentiva a formulação de regras de boas práticas e governança por agentes de tratamento de dados. Essa abordagem tem relação com estratégias de autorregulação, pelas quais o ente regulado, seja a nível individual ou setorial, impõe sobre si mesmo comandos e consequências a esses comandos com vistas a garantir o cumprimento normativo (COGLIANESE e MENDELSON, 2010, p. 3), apesar de a LGPD não trazer, de forma explícita, que o agente de tratamento deva incluir nesse quadro normativo mecanismos sancionatórios.

Vale ressaltar que o §3º do mencionado art. 50, ao ditar que "[a]s regras de boas práticas e de governança deverão ser publicadas e atualizadas periodicamente e poderão ser reconhecidas e divulgadas pela autoridade nacional", também pode ser interpretado de forma a se relacionar à abordagem da teoria da regulação responsiva para a estratégia de autorregulação forçada. De acordo com Aranha (2019), a autorregulação forçada consistiria em exigir do regulado a internalização dos custos de fiscalização por meio da criação de um departamento ou grupo de compliance interno para monitorar o cumprimento das normas e recomendar ações disciplinares contra os infratores.

Se levarmos essas considerações para a aplicação do artigo 20, LGPD, que, como observamos, obriga o controlador de dados a fornecer informações significativas sobre um sistema de decisões automatizadas, desde que não afete os segredos industriais e comerciais, aplicar a regulação responsiva pode ser de grande ajuda para promover maior inteligibilidade de sistemas de ML.

Seguindo nossas considerações anteriores, sistemas de ML são complexos e suas lógicas variam profundamente de uma aplicação para outra. Com frequência, são caixas pretas, o que torna o desafio de compreendê-los cada vez mais complexo. Consequentemente, os reguladores, e especialmente a ANPD, estão frequentemente em uma posição de profunda assimetria de informação para avaliar se as decisões desses sistemas estão sendo responsáveis por práticas ilegais. À medida que a aplicação da LGPD se torna mais madura com o passar dos anos, essa situação pode se tornar gradualmente mais frequente.

Cooperação, a primeira premissa que apresentamos sobre a teoria responsiva, seria uma ferramenta útil neste cenário. Ao convocar os controladores de dados para um diálogo com o objetivo de melhor compreender seus sistemas, a ANPD e outras autoridades regulatórias públicas poderão identificar se, como e em que grau seus sistemas devem ser explicados.

Talvez o primeiro passo para o regulador estabelecer obrigações de inteligibilidade seja entender os riscos colocados por diferentes sistemas e quais informações são úteis em relação ao seu funcionamento para permitir que os titulares dos dados exerçam seus direitos. Cabe aqui buscar paralelos sobre esse assunto em estudos interpretando o RGPD.

Kaminsky e Malgieri argumentam que o RGPD propõe um regime de governança sistêmica por meio do qual os controladores e reguladores de dados estabeleceriam salvaguardas adequadas para o processamento de dados pessoais, inclusive por meios automatizados como em sistemas de ML, por meio de conversas contínuas (KAMINSKY e MALGIERI, 2019, p. 7). Isso é expresso no artigo 40 do RGPD, que prevê a adoção de códigos de conduta e padrões para todo o setor, e refletido no artigo 50 da LGPD, que tem redação muito semelhante.

No que diz respeito ao processamento de dados pessoais por meios automatizados que representam perigos para os titulares dos dados por meio de potenciais decisões tendenciosas, as Diretrizes do Grupo de Trabalho do Artigo 29 (A29WP) sugerem que as empresas verifiquem regularmente seus conjuntos de dados para identificar viéses discriminatórios, e também que revisem regularmente a precisão e relevância das decisões tomadas por seus algoritmos (A29WP, 2017, p. 28).

O regime de governança sistêmica do RGPD é fundamental para permitir que os reguladores e controladores avaliem o grau de fornecimento de informações necessário para cada sistema. Interpretando como o modelo de governança proposto pelo RGPD refletiria na explicabilidade de sistemas de ML, Kaminsky e Malgieri primeiro argumentam que o RGPD propõe um sistema de explicabilidade em múltiplas camadas (*multi-layered explanations*), em que indivíduos têm direito tanto a explicações mais gerais a respeito da lógica de um algoritmo (arts. 13, 14, 15, RGPD) quanto a informações mais específicas sobre decisões individuais tomadas pelo sistema (KAMINSKY e MALGIERI, 2019, p. 5). Nesse sentido, quanto mais intrusivo ou arriscado for um sistema, maiores informações ele deverá divulgar para que os indivíduos possam exercer efetivamente seus direitos. Como concluímos anteriormente, o exercício dos direitos dos titulares dos dados no LGPD também pode exigir que os controladores divulguem informações tanto sobre a lógica geral de um sistema quanto sobre decisões individuais.

Relatórios de impacto de proteção de dados (RIPDs), mencionados acima, e, mais especificamente, avaliações de impacto algorítmico (AIA), teriam um papel determinante em permitir tais explicações. Semelhante aos RIPDs propostos pelo RGPD e pela LGPD, os AIAs funcionariam como ferramenta para obter responsabilidade algorítmica (*algorithmic accountability*) ao avaliar o impacto dos sistemas de inteligência artificial (incluindo de ML) nos direitos dos indivíduos e grupos (KAMINSKY e MALGIERI, 2019, p. 13). Como os AIAs não são especificamente prescritos nem no RGPD nem no LGPD, os autores apresentam sua ideia inspirados nos RIPDs e, devido a fundamentos distintos inerentes a cada um desses dois instrumentos, uma regulamentação adicional provavelmente ainda teria que ser elaborada tanto na Europa quanto no Brasil para tratar sobre AIAs.

RIPDs, quando aplicados para avaliar os sistemas de tomada de decisão automatizada, funcionam para avaliar o grau de risco que o processamento de dados que conduzem representa para as pessoas singulares. Seriam, assim, uma primeira divulgação de informações sobre o funcionamento do sistema avaliado, em uma forma de autorregulação regulada que permitiria também que, uma vez identificados potenciais riscos nesses sistemas, os controladores de dados elaborem formas concretas de mitigá-los (KAMINSKY e MALGIERI, 2019, p. 16).

Por isso, desempenham um papel crucial em uma dinâmica regulatória responsiva. Caso a ANPD, por exemplo, solicite que um controlador de dados divulgue informações sobre um sistema específico, o RIPD ou o AIA podem ser usados como uma evidência de que o controlador adotou todas as medidas ao seu alcance para mitigar os riscos, e assim evitar a escalada de sanções em uma

pirâmide de fiscalização devido à boa fé do controlador. Além disso, RIPDs e AIAs seriam úteis também para avaliar o grau de risco do sistema.

Finalmente, e agora especificamente em relação à inteligibilidade dos sistemas de ML, Kaminsky e Malgieri defendem que controladores publicizem pelo menos um sumário dos achados de seus RIPDs e AIAs. Esse sumário englobaria uma primeira camada de explicação a respeito desses sistemas, contendo principalmente informações sobre a lógica geral do sistema. Essa camada poderia ser complementada por explicações em nível de grupo, para analisar como um algoritmo pode impactar agrupamentos sociais e locais específicos, ou mesmo em nível individual (KAMINSKY e MALGIERI, 2019, p. 27).

Como resultado, RIPDs e AIAs podem funcionar como uma ferramenta eficaz para permitir que reguladores, especialmente a ANPD, entendam como os sistemas de ML aplicados por controladores de dados afetam os direitos e liberdades de indivíduos e grupos e o grau de risco que eles representam. Tal entendimento é fundamental para avaliar a aderência do controlador à LGPD e ao ordenamento jurídico brasileiro como um todo, por meio da análise se o regulado adotou ações mitigadoras cabíveis. Isso é especialmente importante para identificar se o regulado agiu ou não de boa fé, uma característica fundamental do regulamento responsivo para identificar se é hora de escalar as penalidades na pirâmide regulatória. Em segundo lugar, é importante para identificar se são necessárias mais informações para permitir a compreensão do sistema de ML e, consequentemente, permitir a um indivíduo ou grupo o exercício dos direitos previstos na LGPD.

CONCLUSÃO

Neste artigo, discutimos como a teoria da regulação responsiva pode ser um potencial meio de promover a inteligibilidade de sistemas de aprendizado de máquina aplicando a Lei Geral de Proteção de Dados (LGPD). Para isso, primeiro analisamos como termos como inteligibilidade, compreensibilidade, transparência, interpretabilidade e explicabilidade se relacionam entre si quando fazem referência a essas tecnologias.

Em seguida, avançamos para avaliar como LGPD e RGPD lidam com o fornecimento de informações, por controladores de dados, de sistemas de decisão automatizada, que incluem modelos de aprendizado de máquina. Concluímos que, em ambos os regimes, devem ser fornecidas informações suficientes para garantir o exercício dos direitos dos titulares de dados.

Com base no trabalho de principalmente tecnólogos, mapeamos ainda o que significa uma explicação útil e que tipo de decisão deve ser explicada. Apresentamos também uma série de interrogações a serem realizadas a fim de

identificar quais as principais características que um sistema explicativo deve ter em relação ao contexto que se pretende explicar.

Por fim, descobrimos que a regulação responsiva é uma lente poderosa para abordar a aplicação do LGPD e, mais especificamente, para promover a inteligibilidade do aprendizado de máquina. Identificamos que suas duas premissas principais, cooperação e pirâmide de fiscalização, podem ser muito eficazes para permitir que os reguladores, e principalmente a ANPD, entendam melhor esse tipo de tecnologia e combatam vieses algorítmicos. Relatórios de impacto à proteção de dados e avaliações de impacto algorítmico (RIPDs e AIAs, respectivamente) são ferramentas eficazes para demonstrar a conformidade do regulado e também fornecem uma primeira camada de explicação, a partir da qual a ANPD será capaz de avaliar quais informações devem ser fornecidas em relação a um determinado modelo de ML.

REFERÊNCIAS BIBLIOGRÁFICAS

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.

ARTICLE 29 DATA PROTECTION WORKING PARTY (A29WP). ***Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679***, Adopted on 3 October 2017, As last Revised and Adopted on 6 February 2018, WP251rev.01, 29.

ARYA, V. et al. ***One Explanation Does Not Fit All*****:** A Toolkit and Taxonomy of AI Explainability Techniques, 2019, p. 1. Disponível em: https://arxiv.org/abs/1909.03012. Acesso em 27 de agosto de 2020.

AYRES, I e BRAITHWAITE, J. ***Responsive Regulation:*** Transcending the Deregulation Debate. Oxford University Press, USA, 1992.

BAYAMLIOGLU, E. Contesting Automated Decisions: A View of Transparency Implications. **European Data Protection Law Review,**

Volume 4, Issue 4, 2018, pp. 433 - 446. Disponível em doi: https://doi.org/10.21552/edpl/2018/4/6. Acesso em 27 de agosto de 2020.

BILGIC, M.; MOONEY, R.J. ***Explaining recommendations:*** satisfaction versus promotion. Beyond Personalization Workshop, IUI, vol. 5, 153, 2005.

BLACK, Julia e MURRAY, Andrew. Regulating AI and Machine Learning: Setting the Regulatory Agenda. **European Journal of Law and Technology**, Vol 10, Issue 3, 2019.

BOZDAG, E. Bias in algorithmic filtering and personalization. **Ethics Inf Technol,** n. 15, 23 Jun 2013. Disponível em: DOI: 10.1007/s10676-013-9321-6. Acesso em 13 de dezembro de 2020.

BRAITHWAITE, J. To Punish or Persuade: Enforcement of Coal Mine Safety. **Albany: State University of New York Press**, 1985.

BRAITHWAITE, J. Responsive Regulation and Developing Economies. **World Development**, v. 34, n. 5, p. 884 – 898, 2006.

BUCHER, T. ***If… then: algorithmic power and politics.*** Oxford University Press, New York, 1st edition, 2018.

CARVALHO, D.V.; PEREIRA, E.M.; CARDOSO, J.S. ***Machine Learning Interpretability****:* A Survey on Methods and Metrics. Electronics, n. 8, 832, 2019.

COGLIANESE, Cary; MENDELSON, Evan. Meta-Regulation and Self-Regulation. In: BALDWIN, R.; CAVE, M.; LODGE, M. **The Oxford Handbook of Regulation**. Oxford: Oxford University Press, 2010. p. 146-168.

COLLINGRIDGE, D. ***The social control of technology***. Pinter, 1st ed., 1980.

COWAN, N. ***The magical mystery four:*** How is working memory capacity limited, and why? Curr. Dir. Psychol. Sci., n. 19, 2010, pp. 51–57.

DOSHI-VELEZ, F. et al.. ***Accountability of AI Under the Law****:* The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper, 2017. Available at nrs.harvard.edu/urn-3:HUL.InstRepos:34372584. Accessed on 11 December 2020.

DOSHI-VELEZ, F. e KIM, B. ***Considerations for Evaluation and Generalization Interpretable Machine Learning***. Explainable and

Interpretable Models in Computer Vision and Machine Learning, Springer: Berlin, Germany, 2018; pp. 3–17.

DOSHI-VELEZ, F. e KIM, B. ***Introduction to Interpretable Machine Learning***. Proceedings of the CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision, Salt Lake City, UT, USA, 18 June 2018.

DOSHI-VELEZ, F. e KIM, B. ***Towards a rigorous science of interpretable machine learning***. Disponível em http://arxiv.org/abs/1702.08608. Acesso em 30 de novembro de 2020.

FRIEDMAN, D. Does technology require new law. **Harvard Journal of Law e Public Policy**, v. 25, p. 71, 2001. Available at https://digitalcommons.law.scu.edu/facpubs/22/. Accessed on 7 December 2020.

GARCIA, R.C.C. 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.

GILPIN, H.; BAU D.; YUAN, B. Z.; BAJWA, A.; SPECTER, M.; KAGAL, L. ***Explaining Explanations****:* An Overview of Interpretability of Machine Learning. 2019. Disponível em: arXiv:1806.00069. Acesso em 5 de dezembro de 2020.

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.

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

KAMINSKY, M. E. e MALGIERI, G. Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. **International Data Privacy Law**, 2020, forthcoming., U of Colorado Law Legal Studies

Research Paper No. 19-28, Disponível em: https://ssrn.com/abstract=3456224. Acesso em 5 de maio de 2021.

KOLIEB, J. 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.

MCGOUGH, M. ***How bad is Sacramento's air, exactly?*** Google results appear at odds with reality, some say. Sacramento Bee. 2018 August 7.

MILLER, T. ***Explanation in Artificial Intelligence:*** Insights from the social sciences. Artif. Intell. 2018, 267, 1–38.

MITTELSTADT, B. Auditing for Transparency in Content Personalization Systems. **International Journal of Communication** 10(2016), 4991–5002. Disponível em https://www.ijoc.org/index.php/ijoc/article/view/6267. Acesso em 11 ago 2020.

MOLNAR, C. ***Interpretable Machine Learning***. 2019. Disponível em: https://christophm.github.io/interpretable-ml-book/. Accesso em 30 ago 2020.

RÜPING, S. ***Learning Interpretable Models.*** Ph.D. Thesis, University of Dortmund, Dortmund, Germany, 2006.

RUDIN, C. ***Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.*** Disponível em: https://arxiv.org/abs/1811.10154. Acesso em 28 ago 2020.

SEARCH ENGINES IN COLOMBIA: Legal Review and Study of The Muebles Caqueta Vs. Google Inc Case. **Law, State and Telecommunications Review**, [S. l.], v. 12, n. 2, p. 1–13, 2020. DOI: 10.26512/lstr.v12i2.34688. Disponível em: https://periodicos.unb.br/index.php/RDET/article/view/34688. Acesso em: 4 apr. 2021.

SELBST, A. D. e POWLES, J. Meaningful Information and the Right to Explanation. **International Data Privacy Law**, vol. 7(4), 2017, pp. 233-242. Available at SSRN: https://ssrn.com/abstract=3039125. Accessed on 26 August 2020.

STORINO, F.; SENNE, F.; PORTILHO, L.; BARBOSA, A. Unequal Inclusion: An Analysis of the Trajectory of Inequalities in Access, Use and Appropriation of the Internet in Brazil. **Law, State and Telecommunications Review**, *[S. l.]*, v. 12, n. 2, p. 187–211, 2020. DOI:

10.26512/lstr.v12i2.34718. Disponível em: https://periodicos.unb.br/index.php/RDET/article/view/34718. Acesso em: 4 apr. 2021.

UK GOVERNMENT. House of Lords. ***AI in the UK:*** Ready, Willing and Able? 2017. Disponível em: https: //publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10007.htm. Acesso em 21 de novembro de 2020.

WACHTER, S; MITTELSTADT, B; FLORIDI, L. ***Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation***. IDPL, n. 76, 2017.

WEXLER, R. ***When a Computer Program Keeps You in Jail****:* How Computers are Harming Criminal Justice. New York Times. 13 de junho de 2017.

Journal of Law and Regulation
Revista de Direito Setorial e Regulatório

Contact:
Universidade de Brasília - Faculdade de Direito - Núcleo de Direito Setorial e Regulatório
Campus Universitário de Brasília
Brasília, DF, CEP 70919-970
Caixa Postal 04413

Phone: +55(61)3107-2683/2688

E-mail: ndsr@unb.br

Submissions are welcome at: https://periodicos.unb.br/index.php/RDSR