

RACIONALIDADE ALGORÍTMICA, CULTURA DO DESENVOLVIMENTO DE IA/ML E INTRATABILIDADE JURÍDICA DO VIÉS

Sérgio Fernando Ferreira de Lima¹

RESUMO

O presente ensaio investiga por que o viés algorítmico constitui um problema jurídico de difícil ou mesmo impossível resolução nos marcos do direito vigente. A hipótese central é que esta intratabilidade não é acidental, mas estruturalmente condicionada pela articulação de duas racionalidades sobrepostas: a *racionalidade algorítmica* de fundo neoliberal, marcada pela otimização, eficiência e indiferença aos conteúdos substantivos (Aradau; Blanke, 2022; Rouvroy; Berns, 2013; Foucault, 2004) e a cultura epistêmica específica do desenvolvimento de sistemas de aprendizado de máquina (ML/IA), caracterizada pelo que Roberge e Castelle (2021) denominam autorreferencialidade técnica. O argumento percorre três movimentos: (i) a genealogia da racionalidade algorítmica como razão neoliberal aplicada a dados; (ii) a sociologia da produção de ML como campo cultural dotado de lógica própria e déficit de ciências sociais; e (iii) as consequências jurídicas desta dupla estrutura para a regulação do viés, incluindo a crítica aos teoremas da impossibilidade de *fairness* (Chouldechova, 2017; Kleinberg; Mullainathan; Raghavan, 2017) e à insuficiência do modelo de risco adotado pelo EU AI Act e pela LGPD. Propõe-se a conclusão, de que a ordem jurídica brasileira, estruturada em torno de racionalidade substantiva e imputação individual de responsabilidade, encontra diante de si um objeto que opera exclusivamente mediante racionalidade formal (Weber, 1978; Nishant; Schneckenberg; Ravishankar 2024), gerando um hiato estrutural entre o objeto regulado e os instrumentos regulatórios disponíveis.

Palavras-chave: racionalidade algorítmica; viés em IA/ML; regulação algorítmica; intratabilidade jurídica; racionalidade formal/substantiva; neoliberalismo digital.

ABSTRACT

This essay investigates why algorithmic bias constitutes a legal problem that is difficult, or perhaps even impossible, to resolve within the framework of existing law. The central hypothesis is that this intractability is not accidental but structurally conditioned by the convergence of two overlapping rationalities: a neoliberal algorithmic rationality, characterized by optimization, efficiency, and an indifference to substantive content (Aradau; Blanke, 2022; Rouvroy; Berns, 2013; Foucault, 2004), and the specific epistemic culture of machine learning (ML/AI) development, defined by what Roberge and Castelle (2021) term "technical self-referentiality." The argument unfolds in three movements: (i) a genealogy of algorithmic rationality as neoliberal reason applied to data; (ii) a sociology of ML production as a cultural field endowed with its own logic and a deficit of social sciences; and (iii) the legal consequences of this double structure for bias regulation, including a critique of fairness impossibility theorems (Chouldechova, 2017; Kleinberg; Mullainathan; Raghavan, 2017) and the insufficiency of the risk-based model adopted by the EU AI Act and the LGPD. The essay

¹ Doutorando em Direito pela Universidade Federal do Paraná. sergioffernando@gmail.com.

concludes that the Brazilian legal order, structured around substantive rationality and individual attribution of liability, faces an object that operates exclusively through formal rationality (Weber, 1978; Nishant; Schneckenberg; Ravishankar, 2024), resulting in a structural gap between the regulated object and the available regulatory instruments.

Keywords: algorithmic rationality; AI/ML bias; algorithmic regulation; legal intractability; formal/substantive rationality; digital neoliberalism.

1. Introdução: o problema da invisibilidade estrutural do viés

Existe uma tendência recorrente nos debates jurídicos sobre inteligência artificial de tratar o viés algorítmico como um problema técnico contingente, um defeito corrigível por auditoria, uma falha de design superável por transparência, uma anomalia eliminável por regulação de risco. Esta tendência é sintomática de um equívoco mais profundo: a suposição de que os sistemas de aprendizado de máquina (ML) são instrumentos neutros que podem ser avaliados, corrigidos e responsabilizados pelos mesmos critérios aplicáveis a decisões humanas. O presente ensaio recusa esta premissa e propõe que o viés nos sistemas de Inteligência Artificial (IA), na sua atual formulação como Aprendizado de Máquina (*Machine Learning* - ML) é, em larga medida, juridicamente intratável, não porque o Direito seja impotente em geral, mas porque a estrutura de racionalidade que sustenta estes sistemas é fundamentalmente incompatível com a racionalidade substantiva sobre a qual o direito moderno foi construído.

A intratabilidade, aqui, é entendida em sentido preciso: não como impossibilidade absoluta de qualquer resposta jurídica, mas como a condição pela qual os instrumentos jurídicos disponíveis, responsabilidade civil, proibições de discriminação, garantias procedimentais, direitos à explicação, não alcançam as causas estruturais do viés, atingindo apenas suas manifestações superficiais. Esta condição tem uma dupla raiz: de um lado, a racionalidade algorítmica como forma histórica de razão governamental enraizada no neoliberalismo (Aradau; Blanke, 2022, p. 3–7); de outro, a cultura epistêmica peculiar do desenvolvimento de ML, que reproduz internamente esta racionalidade e a torna invisível como escolha normativa (Roberge; Castelle, 2021, p. 4–8).

A ideia aqui exposta é desenvolvida em cinco movimentos. Na seção 2, examina-se a genealogia da racionalidade algorítmica como forma de razão neoliberal aplicada a dados, mobilizando os trabalhos de Aradau e Blanke (2022), Rouvroy e Berns (2013) e a releitura foucaultiana do neoliberalismo como programa de governo por otimização (Foucault, 2004, p. 222–224). Na seção 3, analisa-se a sociologia do desenvolvimento de ML como campo cultural

com lógica própria, a partir da proposta de *sociologia de ponta a ponta* de Roberge e Castelle (2021) e da epistemologia da eficiência descrita por Castelle (2017). A seção 4 articula os dois problemas com a dimensão jurídica, explorando como a distinção weberiana entre racionalidade formal e substantiva (Weber 1978; Nishant; Schneckenberg; Ravishankar, 2024) ilumina o hiato entre os sistemas de ML e os instrumentos do direito. A seção 5 critica os modelos regulatórios vigentes, EU AI Act, LGPD, PL 2.338/2023, como respostas que, ao adotarem o paradigma de risco, consolidam a intratabilidade em vez de superá-la.

2. A racionalidade algorítmica como razão neoliberal: governamentalidade, decomposição e ausência de sujeito

2.1 A governamentalidade sem sujeito

A contribuição mais influente para a compreensão da racionalidade algorítmica como fenômeno político-epistemológico vem da releitura foucaultiana empreendida por Rouvroy e Berns (2013, p. 163–170). Para Rouvroy, a governamentalidade algorítmica configura uma forma radicalmente nova de exercício do poder, ao contrário das tecnologias disciplinares e biopolíticas descritas por Foucault (2004), que operavam sobre corpos e populações por meio da produção de sujeitos normalizados, a governamentalidade algorítmica prescinde do sujeito. Ela atua diretamente sobre os dados, traços de comportamento, padrões de consumo, registros digitais, sem precisar construir uma imagem coerente do indivíduo e sem exigir sua interpretação ou compreensão (Rouvroy; Berns 2013, p. 172–178).

Aradau e Blanke (2022, p. 10–15) sistematizam esta intuição em uma teoria da razão algorítmica² como racionalidade de governo. Para eles, a razão algorítmica não é simplesmente a aplicação de algoritmos a problemas, é uma racionalidade distinta que torna possíveis as práticas de governo e a produção de sujeitos dataficados através da promessa de conhecimento mais preciso e tomada de decisão mais eficiente. Trata-se de uma forma de razão que opera por três movimentos articulados: i) decomposição (desestruturação de objetos complexos em dados elementares), ii) recomposição (reconstituição de perfis, padrões e categorias a partir dos dados)

² O conceito de razão algorítmica (*algorithmic reason*), tal como formulado por Aradau e Blanke (2022), designa uma racionalidade de governo; uma forma histórica de razão que organiza práticas de poder. Ele deve ser distinguido do conceito mais estreito de racionalidade algorítmica como epistemologia interna dos sistemas de ML, analisada por Castelle (2017) e Nishant, Schneckenberg e Ravishankar (2024). Este ensaio opera com ambos os conceitos, distinguindo-os quando necessário.

e iii) particionamento (distribuição de populações em segmentos para intervenção diferenciada) (Aradau; Blanke 2022, p. 45–58). O viés, neste enquadramento, não é uma falha do sistema, é um efeito estrutural do modo como este processo de decomposição-recomposição aplica categorias técnicas a realidades sociais estratificadas.

2.2 Neoliberalismo como epistemologia algorítmica

A conexão entre racionalidade algorítmica e neoliberalismo é mais profunda do que a mera coincidência histórica entre a expansão das tecnologias de ML e a hegemonia neoliberal. Como demonstra Foucault (2004, p. 222–224) em sua análise do neoliberalismo americano (Becker, a Escola de Chicago), a racionalidade neoliberal é fundamentalmente um programa de redução de toda conduta humana a escolhas passíveis de modelagem como algoritmos de otimização de interesse. A leitura econométrica do comportamento humano e a lógica da big data convergem no mesmo gesto, toda ação é sistemática e responde de forma não aleatória às modificações das variáveis do ambiente (Foucault, 2004, p. 269). O indivíduo como capital humano e o indivíduo como perfil de dados são figuras da mesma epistemologia.

Bilić (2018, p. 318–324) demonstra que o crescimento de empresas como o Google se deu a partir de três racionalidades tecnológicas entrelaçadas, organizacional (gestão flexível do trabalho), informacional (extração de valor da audiência como mercadoria) e de acumulação de mais-valia, articulando a tradição da Escola de Frankfurt (Marcuse 1941) com a economia política dos algoritmos. A racionalidade tecnológica marcusiana aplicada ao capitalismo algorítmico contemporâneo revela que a irracionalidade das consequências discriminatórias dos sistemas de ML é produto direto da hiper-racionalidade formal de sua arquitetura, quanto mais eficientemente o sistema otimiza uma função objetivo definida em termos de eficiência de mercado, mais plenamente reproduz as assimetrias do mundo social sobre o qual foi treinado (Bilić, 2018, p. 326).

O ponto crucial, para os fins deste ensaio, é que a racionalidade neoliberal não é apenas o contexto socioeconômico em que os algoritmos são desenvolvidos e implantados, ela é a gramática interna do desenvolvimento técnico de ML. A escolha de uma função de otimização, a minimização de erros, a maximização de métricas de acurácia e o desinteresse pela interpretabilidade são, ao mesmo tempo, escolhas técnicas e normativas de fundo neoliberal (Crawford, 2021, p. 133–138). É esta superposição que torna o viés estruturalmente opaco à problematização jurídica.

3. A sociologia da produção de ML: déficit epistêmico, autorreferencialidade e racionalidade da eficiência

3.1 *A sociologia de ponta a ponta e o déficit de ciências sociais*

Roberge e Castelle (2021, p. 4–8) propõem o que chamam de *sociologia de ponta a ponta* (*end-to-end sociology*) do ML contemporâneo; uma abordagem que atravessasse todo o processo sociotécnico e político da aprendizagem de máquina, da gênese histórica ao impacto social e de volta, em vez de analisar apenas as extremidades do processo (design técnico de um lado, impacto social de outro). Esta proposta responde a uma limitação específica do debate sobre *fairness* em IA, a tendência de tratar os problemas de viés como questões de engenharia *downstream*, passíveis de correção técnica, sem interrogar as condições *upstream* de produção do saber de ML (Roberge; Castelle 2021, p. 12–17). Powles e Nissenbaum (2018), advertem contra a sedutora diversão de resolver o viés como problema de engenharia sem considerar como a noção técnica de *fairness* é implantada na prática.

A noção de déficit de ciências sociais da IA (Sloane; Moss, 2019) aponta para uma característica estrutural do campo, os pesquisadores e praticantes de ML tendem a ver seu próprio comportamento em termos da epistemologia de suas técnicas. Há uma autorreferencialidade que impede a reflexividade crítica sobre as escolhas normativas inerentes no design dos sistemas (Roberge; Castelle 2021, p. 17–19). O engenheiro de ML que define uma função de perda, escolhe um *dataset* de treinamento ou decide sobre um *threshold* de classificação não percebe estas escolhas como decisões políticas, percebe-as como decisões técnicas sobre eficiência de otimização. Esta cegueira reflexiva não é individual, é estrutural, produzida pela cultura epistêmica do campo.

3.2 *A racionalidade da eficiência*

Castelle (2017, p. 6–12), em seu estudo etnográfico com cientistas de dados, identifica uma forma específica de racionalidade epistêmica que rege a avaliação de algoritmos neste campo, não a verdade ou a falsidade (como na lógica dedutiva), nem a correspondência entre modelos matemáticos e estruturas físicas (como na matemática aplicada clássica), mas a *eficiência*. Os algoritmos funcionam bem ou não funcionam (Castelle 2017, p. 6). A avaliação de um sistema de ML prescinde de demonstração lógica e de prova; ela se dá pela performance em métricas definidas operacionalmente. Esta epistemologia da eficiência é a tradução técnica

da racionalidade formal weberiana, o algoritmo é avaliado por sua capacidade de seguir procedimentos predefinidos de otimização, não por sua capacidade de produzir julgamentos substantivamente corretos (Nishant; Schneckenberg; Ravishankar, 2024, p. 21–24).

Este ponto tem consequências jurídicas diretas. A legalidade de uma decisão discriminatória, no direito, não é avaliada por métricas de performance, é avaliada pela presença de motivos ilícitos, pela violação de direitos, pela ausência de justificativa legítima. Um sistema de ML pode ter acurácia de 95% em um *benchmark* e simultaneamente produzir resultados sistematicamente discriminatórios contra minorias raciais, e, do ponto de vista de sua epistemologia interna, continuar sendo considerado um bom sistema. A incomensurabilidade entre a epistemologia técnica do ML e a epistemologia jurídica é, ela mesma, uma fonte estrutural de intratabilidade (Barocas; Selbst, 2016, p. 676–680).

3.3 A sociologia das desigualdades algorítmicas

Shetakofsky (2021) propõe uma agenda de pesquisa que articula a análise das mudanças estruturais produzidas pela IA com a atenção às desigualdades que ela mobiliza e reconstitui. Esta perspectiva complementa a análise de Roberge e Castelle (2021) ao destacar que o déficit de ciências sociais do campo tem consequências materiais distribuídas de forma profundamente desigual, são as populações já marginalizadas que sofrem desproporcionalmente os efeitos do viés, enquanto os benefícios da eficiência algorítmica se concentram em outros grupos (Shetakofsky, 2021, p. 4–7). Noble (2018, p. 1–18) demonstra este padrão empiricamente no caso dos motores de busca, e Buolamwini e Gebru (2018) o quantificam no caso de sistemas de reconhecimento facial, onde as taxas de erro para mulheres negras chegam a ser até 34,7% superiores às de homens brancos.

Para a finalidade da análise jurídica, é relevante notar que esta distribuição desigual dos danos não é capturada pelos instrumentos de responsabilidade individual típicos do direito civil brasileiro. O art. 20 da LGPD e seu equivalente regulatório no PL 2.338/2023 preveem direitos de revisão de decisões automatizadas, mas estes direitos são exercidos individualmente, pressupõem a identificação de um dano específico e imputável, e não alcançam a dimensão coletiva e estrutural da desigualdade produzida algorítmicamente (Doneda, 2019, p. 563–567; Schreiber 2014, p. 98–104).

4. O hiato entre racionalidade formal e racionalidade substantiva: consequências jurídicas

4.1 O paradoxo da racionalização

A distinção weberiana³ entre racionalidade formal e racionalidade substantiva oferece um instrumento analítico interessante para articular a dimensão jurídica do problema. Para Weber (1978, p. 85–86), a racionalidade formal refere-se a cálculos quantitativos e conjuntos rígidos de regras que governam a ação; a racionalidade substantiva refere-se a conjuntos de valores contextualmente acoplados que orientam a decisão. O capitalismo moderno e o direito racional moderno são, ambos, produtos do processo de racionalização formal, e ambos sofrem das mesmas aporias, a irracionalidade das consequências substantivas produzidas por procedimentos formalmente impecáveis.

Nishant, Schneckenberg e Ravishankar (2024, p. 19–25) aplicam esta distinção diretamente ao problema do viés em ML, os algoritmos de aprendizado de máquina operam exclusivamente mediante racionalidade formal, procedimentos matemáticos de otimização que avaliam dados mecanicamente, sem capacidade de exibir bom senso na interpretação de contextos. O viés emerge precisamente desta limitação, sistemas que operam por racionalidade formal herdam e amplificam os vieses presentes nos dados de treinamento, que refletem a racionalidade substantiva historicamente dominante em determinada sociedade, sem a capacidade de reconhecê-los como vieses (Nishant; Schneckenberg; Ravishankar, 2024, p. 28–33).

O paradoxo, do ponto de vista jurídico, é que o direito antidiscriminatório moderno, tanto o internacional quanto o brasileiro, foi construído sobre a pressuposição de que os decisores possuem racionalidade substantiva, a capacidade de compreender valores, de interpretar contextos, de ser responsabilizados por motivos. Quando o decisor é um sistema de ML, que opera exclusivamente por racionalidade formal, o arcabouço conceitual do direito antidiscriminatório perde sua ancoragem. Como Wachter, Mittelstadt e Russell (2021, p. 3–8) demonstram, mesmo o direito europeu de antidiscriminatório, tecnicamente mais elaborado, não foi construído para lidar com a discriminação por *proxy* algorítmico, o mecanismo pelo

³ A escolha de Weber como chave analítica se deve a articulação entre racionalização formal (direito, burocracia, mercado) e as consequências substantivamente irracionais desta racionalização. A aplicação da distinção weberiana ao ML contemporâneo por Nishant, Schneckenberg e Ravishankar (2024) tem, portanto, uma pertinência histórico-teórica que transcende seu uso analítico imediato.

qual sistemas de ML reproduzem discriminação por meio de variáveis correlacionadas com atributos protegidos, sem utilizá-los diretamente (Barocas; Selbst, 2016, 691–698).

4.2 Impossibilidades e a intratabilidade matemática

A intratabilidade do viés em ML apresenta uma dimensão matemática formal, demonstrada pelos chamados teoremas de impossibilidade de *fairness*. Chouldechova (2017, p. 153–158) demonstrou que, para classificadores binários aplicados a grupos populacionais com taxas de base diferentes, é impossível satisfazer simultaneamente calibração por grupos e igualdade de taxas de falsos positivos e falsos negativos. Kleinberg, Mullainathan e Raghavan (2017, p. 1–12) demonstraram resultados análogos em contextos de pontuação de risco⁴. Barocas e Selbst (2016, p. 701–710) identificaram como as práticas padrão de mineração de dados inevitavelmente reproduzem discriminação indireta mesmo na ausência de uso explícito de atributos protegidos.

Estes resultados são juridicamente relevantes porque demonstram que as exigências de *fairness* que o direito poderia razoavelmente impor a um sistema de ML são, em determinadas condições estruturais, matematicamente incompatíveis entre si. Um sistema de pontuação de risco criminal não pode, ao mesmo tempo, ser igualmente calibrado para brancos e negros e ter taxas iguais de falsos positivos para os dois grupos; a menos que as taxas de reincidência sejam idênticas nas duas populações, o que não é o caso quando as condições socioeconômicas diferem estruturalmente (Chouldechova, 2017, p. 160–162). O sistema jurídico que exige ambas as propriedades simultaneamente exige o impossível; o direito que escolhe uma delas em detrimento da outra opta implicitamente por um arranjo distributivo com consequências normativas distintas, sem que esta escolha seja tornada explícita.

4.3 Os limites da ética algorítmica

Diakopoulos (2015, p. 401–408) demonstra que a investigação eficaz de sistemas algorítmicos requer habilidades computacionais específicas, acesso a dados e processos que raramente estão disponíveis, e enfrenta obstáculos legais, como leis de propriedade intelectual e termos de uso, que protegem a opacidade corporativa. Estes obstáculos práticos à

⁴ Os teoremas de impossibilidade de Chouldechova (2017) e Kleinberg, Mullainathan e Raghavan (2017) são resultados matemáticos rigorosos, não meramente afirmações normativas. Sua relevância jurídica reside em demonstrar que, em determinadas condições estruturais (grupos com taxas de base diferentes), as exigências de justiça que o direito poderia razoavelmente impor são matematicamente incompatíveis. Isto não significa que toda regulação do viés é impossível: significa que a regulação precisa escolher explicitamente entre concepções de justiça que não são simultaneamente satisfazíveis, tornando a escolha uma decisão normativa, não técnica.

accountability são amplificados, no nível estrutural, pelas críticas de Aradau e Blanke (2022, p. 153–168) à etificação dos algoritmos como forma de supressão de dissenso.

Para Aradau e Blanke (2022, 160–165), as iniciativas de ética algorítmica incluindo diretrizes, princípios e *frameworks* éticos de grandes empresas tecnológicas, organizações intergovernamentais e sociedade civil, funcionam como tecnologia de governo que visa conduzir a conduta de desenvolvedores e engenheiros, e não como espaço de contestação política substantiva. A etificação dos algoritmos esvazia o dissenso ao focar nas categorias de sujeitos que são interpeladas como sujeitos éticos (os técnicos) e invisibilizando os corpos racializados e generificados mais afetados pelas operações algorítmicas (Aradau; Blanke, 2022, p. 162–163; Crawford, 2021, p. 211–215)⁵.

5. Os limites dos modelos regulatórios vigentes: risco, responsabilidade e a perpetuação da intratabilidade

5.1 O paradigma de risco como legitimação da racionalidade algorítmica

O EU AI Act, o RGPD e em menor grau, a LGPD e o PL 2.338/2023 adotam, como eixo organizador da regulação de IA, o paradigma de risco, os sistemas são classificados por nível de risco (inaceitável, alto, limitado, mínimo) e submetidos a exigências proporcionais. Esta abordagem tem uma virtude pragmática evidente, permite graduar a intensidade regulatória e concentrar recursos de conformidade onde os danos são mais graves. Mas tem um defeito estrutural revelado pela análise anterior, ao aceitar o paradigma de risco, a regulação adota a mesma gramática probabilística dos sistemas que pretende regular (Aradau; Blanke, 2022, p. 198–204).

A regulação por risco pressupõe que os danos são identificáveis *ex ante*, que os *trade-offs* entre eficiência e *fairness* são mensuráveis, e que a conformidade pode ser verificada por auditoria técnica. Nenhuma destas pressuposições é sustentável à luz das análises anteriores; os teoremas de impossibilidade demonstram que não há solução técnica que satisfaça todas as exigências de *fairness* simultaneamente (Chouldechova, 2017; Kleinberg; Mullainathan;

⁵ A crítica à etificação dos algoritmos em Aradau e Blanke (2022) deve ser lida com cuidado para evitar o equívoco oposto: a recusa de qualquer engajamento com a ética algorítmica. O ponto dos autores não é que a ética seja inútil, mas que a ética como código de conduta técnico-corporativo, divorciada da política e da contestação substantiva, funciona como supressão de dissenso. Uma ética da IA politicamente articulada, que tome partido explicitamente sobre as escolhas de design, seria compatível com a abordagem aqui desenvolvida.

Raghavan, 2017); a epistemologia da eficiência do ML torna invisíveis as escolhas normativas embutidas no *design* (Castelle, 2017, p. 8–11); e a racionalidade algorítmica como forma de governamentalidade produz efeitos que se distribuem sistemicamente muito além dos casos individuais auditáveis (Rouvroy; Berns 2013, p. 185–190).

5.2 A LGPD e o art. 20: o sujeito de direito diante do sistema sem sujeito

O art. 20 da LGPD prevê o direito à revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem interesses do titular, bem como o direito a explicações sobre os critérios e procedimentos utilizados. O PL 2.338/2023 avança na direção de exigências de explicabilidade e de avaliação de impacto de direitos fundamentais para sistemas de alto risco. Estas normas são avanços significativos, mas estruturalmente insuficientes.

A insuficiência tem duas dimensões. A primeira é individual-procedimental, os direitos previstos são exercidos individualmente, por titulares específicos, diante de decisões específicas. A dimensão coletiva e estatística do viés algorítmico, que afeta grupos e populações de forma sistemática, não necessariamente por meio de decisões discretas identificáveis, não é capturada por instrumentos pensados para a proteção de interesses individuais (Doneda, 2019, p. 563–567; Tepedino; Teffé, 2019, p. 234–240). A segunda dimensão é epistêmica, a exigência de explicação pressupõe que o sistema é capaz de fornecer uma razão substantivamente compreensível para sua decisão. Mas um sistema de ML que opera por racionalidade formal não tem razões, tem parâmetros, pesos, funções de ativação. A explicação produzida por técnicas de *explainability* (LIME, SHAP etc.) é uma reconstrução *post hoc* que não corresponde ao processo de decisão interno do sistema (Wachter, Mittelstadt; Russell, 2021, p. 18–24; Nishant; Schneckenberg; Ravishankar, 2024, p. 35–37).

5.3 Para além do risco

A crítica aqui desenvolvida não implica que a regulação seja impossível ou inútil. Implica, antes, que a regulação eficaz do viés em ML requer reconhecer o déficit estrutural entre a racionalidade formal dos sistemas e a racionalidade substantiva do Direito, e construir instrumentos que operem nesta lacuna.

A regulação deve alcançar o *upstream* do desenvolvimento, os dados de treinamento, as funções objetivas, os *benchmarks* de avaliação, as decisões de design e não apenas o

downstream da implantação. Roberge e Castelle (2021, p. 16–19) demonstram que é no processo de gênese sociotécnica do ML que as escolhas normativas são inseridas de forma mais profunda e menos visível. Uma regulação que começa apenas na porta de saída dos sistemas já chegou tarde.

Os instrumentos de responsabilidade precisam ser adaptados para capturar a dimensão coletiva, estrutural e transversal do viés. A responsabilidade civil individual e a proteção de dados individuais são insuficientes, nas suas atuais formulações. A regulação precisa politizar explicitamente o que a racionalidade algorítmica de fundo neoliberal naturaliza. A escolha de uma função objetiva de otimização, a definição de um *threshold* de classificação, a seleção de métricas de avaliação são decisões políticas que afetam a distribuição de danos e benefícios entre grupos sociais (Crawford, 2021, p. 210–218). O Direito, como campo de práticas e disputas sobre valores substantivos, pode recusar a naturalização destas escolhas e exigir sua justificação pública, introduzindo, assim, um elemento de racionalidade substantiva no processo de *design* dos sistemas. Neste ponto, a crítica de Aradau e Blanke (2022, 170–175) à ética algorítmica como supressão de dissenso aponta, inversamente, para a condição de uma ética política dos algoritmos, uma ética que não esvazia o dissenso, mas o produz e o organiza.

6. Considerações finais: direito, razão algorítmica e o imperativo da reflexividade

O argumento desenvolvido neste ensaio pode ser condensado em uma proposição: o viés em sistemas de aprendizado de máquina é estruturalmente complexo para o Direito devido a sua incapacidade de uma modo geral de regular novas tecnologias, mas porque existe um hiato fundamental entre a racionalidade formal que governa o funcionamento dos sistemas de ML (Weber 1978; Nishant, Schneckenberg; Ravishankar 2024) e a racionalidade substantiva sobre a qual os instrumentos jurídicos de proteção contra discriminação foram construídos. Este hiato tem raízes históricas, a racionalidade algorítmica como forma de razão neoliberal (Aradau; Blanke 2022; Foucault 2004), culturais, a epistemologia da eficiência como horizonte exclusivo de avaliação no desenvolvimento de ML (Roberge; Castelle, 2021; Castelle 2017), e matemáticas, os teoremas de impossibilidade de *fairness* (Chouldechova, 2017; Kleinberg; Mullainathan; Raghavan 2017).

A resposta jurídica adequada a este hiato não pode ser apenas técnica, não pode consistir em exigir mais transparência, mais auditoria, mais conformidade procedimental. Ela precisa ser reflexiva, precisa reconhecer que a regulação do viés é, inescapavelmente, um campo de disputa

sobre valores substantivos, sobre quais formas de igualdade são desejáveis, sobre quais *trade-offs* são aceitáveis, sobre quais populações devem suportar quais riscos. Tornar estas disputas explícitas e politicamente articuladas, em vez de deixá-las obscurecidas pela gramática técnica da otimização algorítmica (Aradau; Blanke, 2022, p. 163; Crawford, 2021, p. 211), é a condição para que o Direito possa contribuir efetivamente para a construção de sistemas de IA menos discriminatórios.

A intratabilidade do viés, ao final, revela menos uma limitação do Direito do que uma tensão profunda entre dois projetos de racionalização que se desenvolveram em paralelo na modernidade, o projeto jurídico de imputação de responsabilidade e proteção de direitos substantivos, e o projeto técnico-econômico de otimização algorítmica (Bilić 2018; Rouvroy; Berns 2013). Fazer o Direito responder adequadamente ao viés em IA é, em alguma medida, retomar a disputa entre estes dois projetos e recusar a rendição do primeiro ao segundo. Neste sentido, o diagnóstico de intratabilidade não é uma sentença de resignação, é o ponto de partida para uma regulação que leve a sério a dimensão política do que a gramática técnica da eficiência insiste em apresentar como neutro.

REFERÊNCIAS

- ARADAU, Claudia, e Tobias BLANKE. **Algorithmic Reason: The New Government of Self and Other**. Oxford: Oxford University Press, 2022
- BAROCAS, Solon, e Andrew D. SELBST. Big Data's Disparate Impact. **California Law Review**, v.104, n. 3, p. 671–732, 2016.
- BILIĆ, Paško. A Critique of the Political Economy of Algorithms: A Brief History of Google's Technological Rationality. **tripleC: Communication, Capitalism & Critique**, v.16, n. , p.315–331, 2018.
- BRASIL. **Lei n. 13.709, de 14 de agosto de 2018** (Lei Geral de Proteção de Dados Pessoais LGPD). Brasília: Presidência da República, 2018.
- BRASIL. **Projeto de Lei n. 2.338, de 2023**. Regula a inteligência artificial no Brasil. Brasília: Senado Federal, 2023.
- BUOLAMWINI, Joy; Timnit, GEBRU. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. **Proceedings of Machine Learning Research**, n. 8, p. 1–15, 2018.
- CASTELLE, Michael. Algorithmic Rationality: Epistemology and Efficiency in the Data Sciences. **Social Studies of Science**, v. 48, n. 1: p.1–22, 2017.

- CHOULDECHOVA, Alexandra. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. **Big Data**, 5 (2): 153–163, 2017.
- COULDRY, Nick; MEJIAS, Ulises A. **The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism**. Stanford: Stanford University Press, 2019.
- CRAWFORD, Kate. **Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence**. New Haven: Yale University Press, 2021.
- DIAKOPOULOS, Nicholas. Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. **Digital Journalism**, v. 3, n. 3, p. 398–415.
- DONEDA, Danilo. A proteção de dados pessoais nas relações de consumo: para além da informação creditícia. In **Lei Geral de Proteção de Dados Pessoais e suas repercussões no Direito Brasileiro**, TEPEDINO, Gustavo et al. (Org.), p. 558–575. São Paulo: Thomson Reuters, 2019.
- FOUCAULT, Michel. **Naissance de la biopolitique: Cours au Collège de France (1978–1979)**. Paris: Gallimard/Seuil, 2004.
- KLEINBERG, Jon, MULLAINATHAN, Sendhil; RAGHAVAN, Manish. Inherent Trade-Offs in the Fair Determination of Risk Scores. **Proceedings of Innovations in Theoretical Computer Science**: 1–23, 2017.
- MARCUSE, Herbert. Some Social Implications of Modern Technology. **Studies in Philosophy and Social Science**, v. 9, n. 3, p. 414–439, 1941.
- NISHANT, Rohit, SCHNECKENBERG, Dirk; RAVISHANKAR, M.N. The Formal Rationality of Artificial Intelligence-Based Algorithms and the Problem of Bias. **Journal of Information Technology**, v. 39, n. 1, p. 19–40, 2024.
- NOBLE, Safiya Umoja. **Algorithms of Oppression: How Search Engines Reinforce Racism**. New York: New York University Press, 2018.
- ROBERGE, Jonathan; CASTELLE, Michael (Org). **The Cultural Life of Machine Learning: An Incursion into Critical AI Studies**. Cham: Palgrave Macmillan, 2021.
- ROUVROY, Antoinette; BERNIS, Thomas. Gouvernamentalité algorithmique et perspectives d'émancipation. **Réseaux**, n. 177, p. 163–196, 2013.
- SCHREIBER, Anderson. **Direitos da Personalidade**. 3. ed. São Paulo: Atlas, 2014.
- SHESTAKOVSKY, Benjamin. Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change. **Socius: Sociological Research for a Dynamic World**, n. 7, p. 1–11, 2021.
- TEPEDINO, Gustavo; Chiara Spadaccini de TEFFÉ. Consentimento e proteção de dados pessoais na LGPD. In **Lei Geral de Proteção de Dados Pessoais e suas repercussões no Direito Brasileiro**, (Org.), p. 226–253. São Paulo: Thomson Reuters, 2019.

WACHTER, Sandra; MITTELSTADT, Brent; RUSSELL, Chris. Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. **Computer Law & Security Review**, n. 41, p.1–24, 2021.

WEBER, Max. **Economy and Society: An Outline of Interpretive Sociology**. Berkeley: University of California Press, 1978.