# Viés algorítmico – um balanço provisório

*Letícia SIMÕES-GOMES*[*]
*Enrico ROBERTO*[**]
*Jônatas MENDONÇA*[***]

**RESUMO:** Este artigo se insere no campo da Sociologia Digital e objetiva realizar um balanço bibliográfico sobre viés algorítmico (*algorithmic bias*) nas Humanidades, observando quais definições, causas, diagnósticos e perspectivas são apresentadas para o fenômeno. Tomando por base artigos presentes nos principais portais de periódicos nas línguas inglesa, espanhola e portuguesa, foram encontrados majoritariamente artigos de cunho ensaístico, produzidos no Norte Global, com baixa penetração da temática nas literaturas em espanhol e português. Percebe-se certa indefinição sobre o termo, ora tratado como sinônimo de discriminação, ora como sua causa. Como principais fontes de viés, foram identificadas a construção de ferramentas e os dados de treinamento. Esses fatores ressaltam a necessidade de aumentar a transparência no desenvolvimento de algoritmos; ademais, sugerem tendência analítica de enfatizar o caráter subjetivo do viés algorítmico. Esses achados salientam a importância de integrar à análise elementos que transcendem a subjetividade desses atores.

**PALAVRAS-CHAVE:** Viés algorítmico. Algoritmo. Discriminação. Sociologia Digital.

[*] USP – Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Programa de Pós-Graduação em Sociologia. Núcleo de Estudos da Violência. São Paulo – SP – Brasil. 05508-080 - le.simoesgomes@gmail.com. https://orcid.org/0000-0002-6329-6115

[**] USP – Universidade de São Paulo. Faculdade de Direito na USP. São Paulo – SP – Brasil. 01005-010 - enrico.roberto@live.com. https://orcid.org/0000-0001-9438-9173

[***] USP – Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Programa de Pós-Graduação em Sociologia. São Paulo – SP – Brasil. 05508-080 - mendonca.jonatas@gmail.com. https://orcid.org/0000-0002-1746-8131

## I. Introdução

A vida social contemporânea está impregnada por novas tecnologias digitais[1] e pelos algoritmos que as compõem, sendo usadas para interação com outras pessoas e instituições, para movimentação por espaços públicos, para a produção e recriação de subjetividades (LUPTON, 2015). Tal processo de normalização das tecnologias digitais no mundo social gera uma infinitude de dados e metadados sobre hábitos, padrões de consumo e de interação. Essas tecnologias se desdobram em uma miríade de produtos – de objetos inteligentes (*smart objects*), conscientemente adquiridos e manipulados, a sensores e outros aparelhos dispersos pelos territórios por onde se circula (como *smart cameras*, ou outros sensores de ambiente). Imperativo para o tratamento e utilização deste *big data*[2] é o uso de sistemas algorítmicos que automatizem os processos de tratamento, categorização e filtragem de dados.

A partir da aplicação de tais modelos algorítmicos a dados digitais[3] é possível extrair correlações entre categorias, que são postas a serviço dos mais diversos propósitos. Um deles é a manutenção dessas tecnologias. Outras aplicações passam pela consolidação e comercialização de dados agregados com objetivo de aumentar a sua precisão sobre agrupamentos sociais com fins de *micro-targeting*, por treinamento de outros sistemas algorítmicos, pela promoção de controle e vigilância estatais, entre outros. Daqui extraem-se três questões: a primeira é que esses dados, automaticamente coletados, tornam-se mercadoria quando agregados. A segunda é que, uma vez vendidos, são utilizados em outras atividades às quais não estão inicial e diretamente relacionados, além de serem manipulados com o auxílio de processos automatizados. Finalmente, ressalta-se a interatividade entre indivíduos, dados e análise, uma vez que tais tecnologias incidem e geram impactos sobre o mundo e a vida social (MARRES, 2017)[4].

Alguns estudiosos apontam que a tendência de valorização e comercialização de dados está associada a uma orientação da lógica capitalista em termos de ganhos em produtividade e rentabilidade (ZUBOFF, 2018; 2019; FOURCADE; HEALY, 2016), além de a processos mais amplos de quantificação da vida social e de movimentos de categorização (PASQUALE, 2017; POWER, 2004; LE GALÈS,

[1] O termo tecnologia digital se refere à conjunção de *software* e *hardware* em um dispositivo, bem como à infraestrutura necessária para seu funcionamento (LUPTON, 2015, p.7).

[2] Boyd e Crawford (2012, p.663) argumentam que a característica primordial do *big data* não é a grande quantidade de dados reunida em uma base, mas a capacidade de buscá-los, agregá-los e cruzá-los. Para elas, *big data* é um fenômeno "cultural, tecnológico e acadêmico" que articula tecnologia, análise e mitologia.

[3] Dados digitais, aqui, dizem respeito ao registro e transmissão de dados codificados (em sequências binárias) por meio de tecnologias digitais (LUPTON, 2015).

[4] Em que medida "interação" permanece sendo um conceito adequado para a descrição e denominação desse fenômeno, foge ao escopo deste artigo.

2016; HACKING, 1990)[5]. Apresentados e vendidos como potenciais soluções a problemas sociais, sistemas algorítmicos são mobilizados como formas de aumentar a eficiência de processos os mais variados (CASTRO, 2018; MACHADO, 2018), de modo mais objetivo e cego a diferenças sociais, colocados como mais neutros, previsíveis e precisos (BOYD; CRAWFORD, 2012; LUPTON, 2015)[6].

Há uma literatura em expansão polemizando tais afirmações, respaldada em campos consolidados de investigação sobre tecnologias, discriminação, desigualdade e quantificação. De matriz multidisciplinar, esses estudos abordam a produção de vieses algorítmicos (*algorithmic bias*), discriminação algorítmica (*algorithmic discrimination*), e filtragem/perfilamento algorítmico (*algorithmic profiling*). Dentro desse contexto, apresenta-se uma revisão bibliográfica acerca da temática do viés algorítmico nas Humanidades, com base nos principais portais de periódicos nas línguas inglesa, espanhola e portuguesa. O objetivo é acompanhar o surgimento e desdobramento do debate sobre o viés algorítmico, com ênfase nas definições, perspectivas e nos diagnósticos que circulam na literatura acadêmica neste momento.

Após introdução aos conceitos centrais da área, a segunda seção descreve a metodologia usada para o levantamento e sistematização dos trabalhos. A terceira seção discute a literatura internacional sobre viés algorítmico, destacando as famílias de definições presentes, as causas e as consequências apontadas para a existência de viés algorítmico, além de quais perspectivas são apresentadas para reduzi-lo ou superá-lo. A quarta seção trata dos trabalhos em português, atentando para a abordagem do tema no Brasil e para eventuais diálogos com perspectivas internacionais. A quinta e a última seções discutem tendências identificadas e tecem considerações para o subsequente desenvolvimento de pesquisas sobre o tema.

Como ponto de partida, cabe definir o que se entende por algoritmo. Em um sentido amplo, um algoritmo é uma sequência de instruções a serem cumpridas em uma determinada ordem, ou "procedimentos codificados que, com base em cálculos específicos, transformam dados em resultados desejados" (GILLESPIE, 2018, p.97). No entanto, é usada aqui a acepção contemporânea de "algoritmos digitais", como "entidades reais que consistem em operações finitas de cálculo, bem como sequências incomputáveis de dados" (DIXON-ROMÁN; NICHOLS; NYAME-MENSAH, 2019, p.4, tradução nossa). Nesse ponto, convém notar que

[5] Assunto que se discute, na sociologia, pelo menos desde a filosofia do dinheiro.

[6] Boyd e Crawford (2012, p.663, tradução nossa) destacam o aspecto "mitológico" do *big data*, isto é, da "crença generalizada de que grandes conjuntos de dados oferecem uma forma superior de inteligência e conhecimento que pode gerar *insights* que antes eram impossíveis, com a aura de verdade, objetividade e precisão".

se referir ao viés algorítmico trata-se do enviesamento estatístico[7] ou moral[8], ou da discriminação causada por meios algorítmicos –, isto é, da automatização de cálculos e processamento automatizado de dados.

Com base na literatura pesquisada realiza-se uma tentativa inicial de apresentar o que se entende por "viés".

Embora haja diversas modalidades de algoritmos, o fenômeno do viés algorítmico é associado primordialmente ao uso de algoritmos de aprendizado de máquina (*machine learning*). Neste processo, o sistema apreende padrões contidos em um conjunto de dados com o qual é alimentado previamente, e faz uso de tais padrões para chegar a resultados não explicitamente programados pelos seus desenvolvedores, além de projetar tais padrões apreendidos a novas situações (ALPAYDIN, 2017).

Pode-se assim afirmar que o aprendizado de máquina é uma forma de processamento de dados, isto é, de taxonomia. Estudos a esse respeito chamam atenção para a construção de classes de equivalência, desde as etapas de seleção das qualidades, constituição das categorias à modelagem, gerando informações que modificam interpretações da realidade, bem como a realidade em si (DESROSIÈRES, 2011). Por exemplo, um algoritmo ao qual são apresentadas inúmeras imagens de rostos anotados, por humanos, com as classificações "homem" ou "mulher", será capaz de identificar padrões distintos (uso de cabelo curto ou comprido, maquiagem etc.) de cada uma das fotos a eles alimentadas e, posteriormente, classificar outras fotos de rostos que não faziam parte do conjunto que o alimentou inicialmente. Segundo Desrosières (2011), essas práticas taxonômicas foram, historicamente, objeto de disputa. Agora, "[o] estabelecimento da concordância [passa a ser] deslocado para a construção negociada das próprias máquinas. Porém, mesmo então, a controvérsia sobre esses mecanismos pode ser novamente desencadeada" (DESROSIÈRES, 2011, p.278, tradução nossa). Seguindo essa visão, a discussão sobre o viés algorítmico é uma das formas de reavivamento dessas controvérsias.

## II. Metodologia

Entre os dias 15 e 30 de janeiro de 2020, foram buscados artigos científicos publicados em periódicos nacionais, latinoamericanos (língua espanhola), e de

[7] O viés estatístico ou técnico deriva da manipulação de classes de generalizações em um banco de dados que, por motivos técnicos, geram resultados desiguais entre grupos, ainda que não tenha como objetivo expresso a discriminação e reprodução de desigualdades.

[8] Viés moral refere-se ao uso de generalizações baseadas em preconceitos visando a resultados discriminatórios. Conforme será discutido, a diferenciação entre viés moral e estatístico, embora analiticamente possível, tem aplicação comprometida ao pressupor que a produção de vieses técnicos estaria desvinculada de considerações morais.

língua inglesa. A opção por esses idiomas baseou-se no reconhecimento de que esta temática surgiu inicialmente em países com implementação mais acelerada de tecnologias digitais em domínios cruciais da vida social – como, por exemplo, a automação de processos relacionados ao provimento de serviços estatais e privados[9] –, países nos quais domina a literatura científica em língua inglesa. Além disso, interessa investigar como essa discussão ressoa em outros centros de produção científica mais próximos ao contexto brasileiro, se há apropriação desta temática, quais questionamentos são trazidos e se há inovações teórico-metodológicas.

Devido à profusão de artigos em língua inglesa, a busca se concentrou em periódicos da área de Humanidades, com foco no campo da Sociologia. Desse modo, uma produção significativa de livros, relatórios, capítulos de coletâneas e anais de congresso foi descartada, a despeito da sua reconhecida relevância para a difusão e discussão internacional do tema[10]. Por ser uma produção recente não foi estabelecido recorte temporal; os bancos de dados selecionados (pela abrangência) podem ser visualizados no Quadro 1.

**Quadro 1 –** Portais de periódicos e bancos de dados de artigos consultados

| Nacionais | Latinoamericanos | Demais países |
|---|---|---|
| BIB[1] | Scielo | Google Scholar |
| Banco de teses da CAPES | Latindex | HeinOnline |
| Anais LAVITS[1*] | | DOAJ[4] |
| Anais Humanidades Digitais[*] | | Taylor & Francis |
| Anais ANPOCS[3*] | | JSTOR |

Fonte: Elaboração própria.

[*] Anais incluídos conforme discussão apresentada adiante
[1] Revista Brasileira de Informação Bibliográfica em Ciências Sociais
[2] Rede Latino-Americana de Estudos sobre Vigilância, Tecnologia e Sociedade
[3] Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais
[4] Directory of Open Access Journals

Em cada um deles, foram procurados artigos com os termos apontados no Quadro 2. Foram utilizadas as palavras-chave que estão na coluna à esquerda, em combinação ou não com os termos da coluna à direita.

[9] Apesar de o caso chinês se enquadrar neste aspecto, os autores não dominam a leitura em mandarim.
[10] Dentre os livros, destaca-se O'Neil (2016), Eubanks (2018), Pasquale (2015), Custers *et al.* (2012), Noble (2018), Benjamin (2019a), Benjamin (2019b).

**Quadro 2 –** Palavras-chave utilizadas para busca dos artigos

| Termo | Combinação |
|---|---|
| viés algorítmico / algorithmic bias / sesgo algorítmico | N/A[1] |
| filtragem algorítmica / algorithmic profiling / perfil algorítmico & perfilado algorítmico | N/A |
| algoritmo / algorithm | Discriminação / discrimination / discriminación |
| inteligência artificial / artificial intelligence / inteligencia artificial | |
| big data | OU |
| data / dados; digital | |
| Software | Viés / bias / sesgo |

Fonte: Elaboração própria

[1]Não aplicável

A depender da base, foram encontrados entre 300 e 10 mil resultados. No caso da busca em inglês pelo Google Scholar, que reúne artigos já contidos em outras bases de dados, restringiu-se excepcionalmente a busca à expressão "viés algorítmico", entre aspas. No JSTOR foram buscadas todas as combinações, porém restritas às entradas em revistas de Sociologia (recurso não disponível no Google Scholar). Os títulos e resumos dos resultados foram considerados manualmente, para assegurar que os artigos selecionados efetivamente abordassem o tema.

Ao cabo do levantamento bibliográfico, restaram apenas cinco artigos em língua portuguesa, o que suscitou a necessidade de busca de fontes alternativas na literatura nacional. Optou-se por agregar teses e dissertações, bem como artigos em anais de congresso. O acesso a essas produções ocorreu de duas formas principais: pela expansão dos critérios de recorte nos portais de busca acadêmica (como o Google Scholar) e pelo acesso direto a Anais de congressos relevantes para o tema: o Simpósio Internacional Lavits, os Encontros da ANPOCS e o Congresso Internacional em Humanidades Digitais[11]. No que diz respeito às teses e dissertações, recorreu-se ao Banco de Teses da CAPES. Com essa medida a amostra de trabalhos nacionais foi ampliada para 13 artigos e nove dissertações.

Para sistematizar a análise do conteúdo de todos e cada um dos trabalhos selecionados, foi utilizado um formulário com 16 tópicos, elencados no Quadro 3. Para comparação entre gêneros textuais semelhantes, as dissertações foram

[11] A escolha pela não inclusão dos Anais dos Congressos da Sociedade Brasileira de Sociologia (SBS) se deu por considerações práticas: a inconstância da existência dos Anais e a falta de um endereço eletrônico único disponível limita a investigação.

destacadas em formulário à parte, ainda que dotado dos mesmos pontos de inquérito. A pré-estruturação de respostas a algumas questões, apesar de constituir um direcionamento do olhar do pesquisador ao seu objeto, possibilita sua comparação para amostras maiores. Além do título do artigo, foram coletados os seguintes dados:

**Quadro 3 –** Tópicos abordados no formulário para coleta de dados do artigo

| **Tópicos** | |
|---|---|
| Número de autores | Definição de viés algorítmico |
| Ano de publicação | Causas do viés algorítmico |
| Nacionalidade do Artigo | Consequências do viés algorítmico? |
| Distribuição geográfica dos autores (pertencimento institucional) | Soluções para o viés algorítmico |
| Subgênero textual | Perspectiva teórico-metodológica |
| Distribuição geográfica do objeto empírico | Hipótese |
| Área do periódico | Principais argumentos |
| Tema do artigo | Palavras-chave[1] |
| Objeto do artigo | Outros comentários sobre uso do termo viés algorítmico |

Fonte: Elaboração própria.

[1] Neste caso, palavras-chave, ou tags, não correspondem necessariamente àquelas propostas pelos autores, mas dizem respeito a tópicos que foram abordados com alguma profundidade ao longo do texto.

Após a elaboração e pré-teste do formulário, os autores padronizaram seu preenchimento. Ao longo da leitura, ficou claro que a pré-seleção de textos continha entradas que tratavam muito subsidiariamente – ou não tratavam em absoluto – de viés algorítmico; essas foram subtraídas da amostra. Eventuais referências derivadas da leitura desse *corpus* de textos foram incorporadas, respeitados os requisitos acima enunciados. Após esse processo, houve 123 entradas ao formulário, sendo 11 delas brasileiras. A este resultado somaram-se sete dissertações.

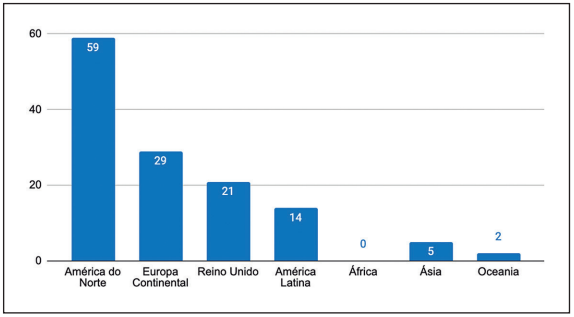
## III. Literatura Internacional

### A. Panorama sobre a produção

Dos 123 artigos analisados, 110 estão em inglês e publicados em revistas de circulação internacional, sem periódicos brasileiros e nem de outros países da América Latina. Para as análises, foram considerados os 123 artigos. A

predominância de artigos publicados por autores do Norte Global é evidente, e ao analisar a procedência dos estudos por continente, constata-se que a América do Norte, a Europa Continental e o Reino Unido somam 89,4% de toda a produção[12]. Em um único artigo não foi possível coletar os dados, pois os autores se apresentam como empresários, e não acadêmicos.

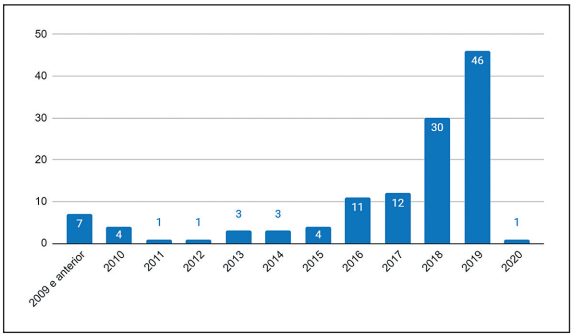
**Gráfico 1 –** Vinculação institucional dos autores, por distribuição geográfica




Fonte: Elaboração própria.

Vale notar a evolução temporal das publicações na área: 46 dos artigos analisados foram publicados em 2019, sendo que 83,4% de toda a produção científica sobre viés algorítmico ocorreu a partir de 2016 (Gráfico 2).

**Gráfico 2 –** Ano de publicação



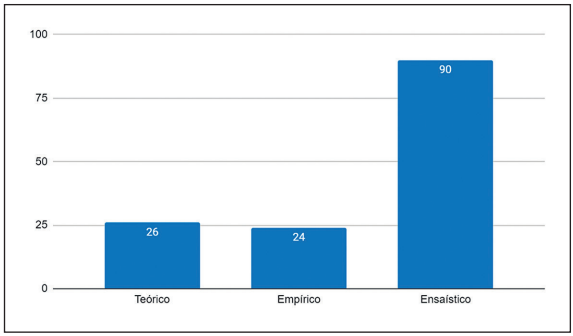

Fonte: Elaboração própria.

Em relação à autoria, 52,8% foram escritos individualmente e 47,2% por dois ou mais autores.

[12] Ressalte-se que o Reino Unido foi contabilizado separadamente da Europa Continental em virtude da quantidade de contribuições e das conhecidas diferenças de tradição em pesquisas nas Humanidades, em especial na Sociologia e no Direito.

Os artigos foram agrupados nas rubricas teóricos, empíricos ou ensaísticos[13], com a possibilidade de que um texto fosse classificado em mais de uma delas. O modo ensaístico predominou, presente em 90 artigos (74,4% do total).

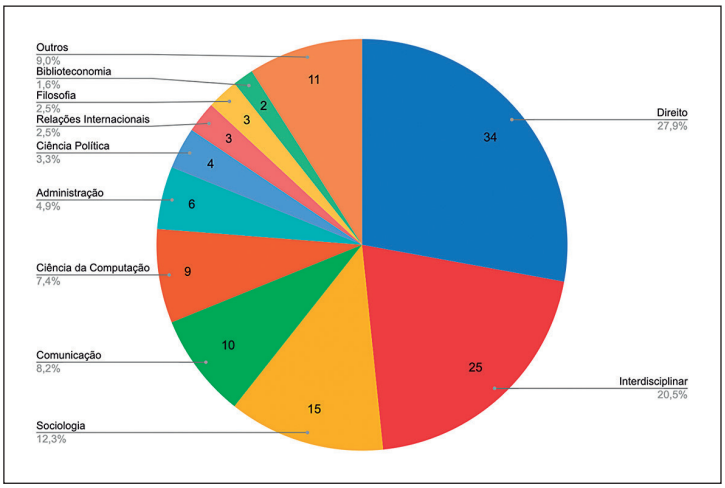
**Gráfico 3 –** Tipo do Artigo




Fonte: Elaboração própria.

Por fim, destaca-se que os textos foram majoritariamente publicados em revistas jurídicas, a seguir em periódicos interdisciplinares e de sociologia (Gráfico 4).

**Gráfico 4 –** Campo das Humanidades do Periódico




Fonte: Elaboração própria.

[13] Teóricos foram considerados os artigos que tratam especificamente de determinada teoria ou propõem adições a teorias existentes; empíricos aqueles que coletam e realizam análise a partir de dados primários; ensaísticos aqueles que não apresentam dados primários nem contribuições teóricas, mas resenham o tema.

Na literatura jurídica, nota-se a abordagem do viés algorítmico no contexto de outro problema – no uso para a persecução penal, com recorrência do caso COMPAS[14] (HUQ, 2019; RICH, 2016); no ambiente de trabalho (KIM, 2017); no contexto da proteção de dados (MANN; MATZNER, 2019); em perfilamento (HILDEBRANDT; KOOPS, 2010); e no direito de autor (BURK, 2019). Em todos os casos prevalece a tentativa de subsumir[15] normas norte-americanas, em especial aquelas relativas a direitos fundamentais individuais de privacidade e de não-discriminação, às realidades impostas pelo fenômeno.

Um segundo aspecto interessante é que boa parte da literatura jurídica busca apresentar e categorizar o fenômeno e suas consequências, além de explorar possíveis soluções normativas sistêmicas, como maior transparência, auditabilidade, responsabilização (*accountability)* etc. (AYRE; CRANER, 2018; BAROCAS; SELBST, 2016; MAYSON, 2018; LA FORS; CUSTERS; KEYMOLEN, 2019; DIAKOPOULOS, 2016). Diakopoulos (2016), por exemplo, propõe pensar *accountability* e transparência com base na divulgação de informações sobre intenções e envolvimento humanos, qualidade dos dados e do modelo, quais as especificidades das inferências realizadas pelo algoritmo, e como e para o que o algoritmo é mobilizado (denominado por ele "presença algorítmica").

A literatura jurídica trata o viés algorítmico praticamente como um sinônimo de discriminação – enquanto fenômeno vivenciado do ponto de vista individual, e causado, também, por agentes individuais (programadores negligentes, juízes com vieses inconscientes etc.). Poucos autores nessa literatura (BAROCAS; SELBST, 2016, AYRE; CRANER, 2018, MANN; MATZNER, 2019) exploram em detalhe as dimensões sistêmicas e propriamente sociais – como o aprofundamento de desigualdades – do uso de vieses algorítmicos.

## B. Conteúdo dos artigos

Em relação à definição de "viés algorítmico", foi surpreendente perceber que *nenhum* dos trabalhos oferece uma definição explícita para o fenômeno. Nota-se na literatura estudada diferentes abordagens e rara preocupação com uma delimitação teórica clara (FAVARETTO; DE CLERCQ; ELGER, 2019).

Em muitos casos, viés algorítmico, viés e discriminação são tratados quase como sinônimos, associados à falta de justeza (*fairness*) ou à violação de outros

[14] Trata-se de *software* usado nos EUA para computar notas de risco de reincidência para réus, as quais são utilizadas pelos juízes do processo no sopesamento da pena. Neste caso, foi indicado que, controladas outras variáveis, o risco de reincidência de negros era invariavelmente mais alto que o de brancos.

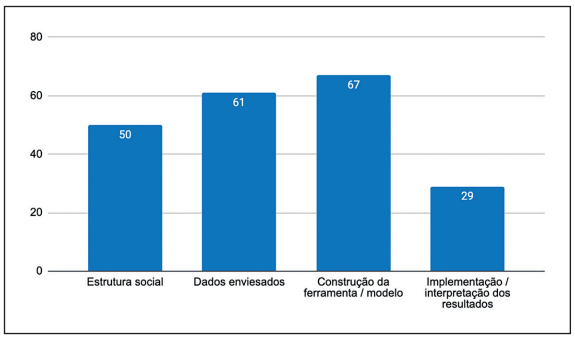
[15] No direito, fala-se de "subsunção" ou "teste de subsunção" quando se busca testar a aplicabilidade de uma norma a um fato concreto.

direitos (HOWARD; BORENSTEIN, 2018, MCGREGOR; MURRAY; NG, 2019; LEPRI *et al.*, 2018; GRAHAM; WOOD, 2003; FRIEDMAN; NISSENBAUM, 1996; MBADIWE, 2018; HILDEBRANDT; KOOPS, 2010). Nesse contexto, Lim e Taeihagh (2019, p.7, tradução nossa) resumem a tendência: "Um sistema é considerado enviesado quando contém características 'intencionais' ou 'não intencionais' que discriminam injustamente certos indivíduos ou grupos de indivíduos na sociedade".

Outros autores definem viés algorítmico como a própria realidade da distorção de dados, algoritmos e resultados de acordo com uma realidade social pré-existente – ora com maior destaque ao desenvolvimento do modelo e ao uso dos dados, ora à própria estrutura social (AYRE; CRANER, 2018; SHIN; FOTIADIS; YU, 2019; BAROCAS; SELBST, 2016).

Buscou-se também conhecer as causas apresentadas para a ocorrência de viés algorítmico. Por se tratar de fenômeno majoritariamente associado ao aprendizado de máquina, procurou-se identificar se o viés decorria (i) do momento em que a máquina era alimentada com dados (*input* ou dados de treinamento), (ii) da construção da ferramenta ou modelo, (iii) da má interpretação do resultado ou *output* do algoritmo, ou (iv) de fatores externos ao sistema. No formulário, as opções não eram excludentes[16].

**Gráfico 5 –** Causas apontadas para o viés algorítmico




Fonte: Elaboração própria.

Nos 98 artigos que mencionam causas para o fenômeno, predominam preocupações com o momento de construção do modelo. Nesse ponto do processo o viés pode emergir na seleção de *proxies* apropriados, na definição do tipo de algoritmo aplicado, no processo de categorização das variáveis, entre outros.

[16] A divisão proposta para estruturar as causas do viés algorítmico na literatura não é unívoca, sendo aqui utilizada no intuito de mapear a argumentação. Taxonomias ricas são propostas, por exemplo, por Friedman e Nissenbaum (1996), Danks e London (2017), Barocas e Selbst (2016), Favaretto, De Clercq e Elger (2019) e Rich (2016).

Eckhouse e colegas (2019, p.205, tradução nossa) fazem uma reflexão muito apropriada sobre o potencial viés na construção de um algoritmo: "O processo de construção desses modelos requer julgamento humano sobre o que significa justiça em termos matemáticos e quando é moralmente aceitável julgar as pessoas com base no comportamento dos outros"[17].

Diversos autores discutem o viés gerado pela coleta de dados enviesados, ou de forma geral, no *input* do sistema. É frequente a referência à reprodução de padrões sociais pré-existentes em virtude da automação baseada em dados passados (AYRE; CRANER, 2018; SHIN; FOTIADIS; YU, 2019; BAROCAS; SELBST, 2016). Da mesma forma, parte da literatura enfatiza a estrutura social subjacente ao viés algorítmico (HILDEBRANDT; KOOPS, 2010; BENJAMIN, 2016; STRAUß, 2018; ATKINSON, 2018; MBADIWE, 2018).

Por fim, quanto à possibilidade de o viés ter origem no uso dos resultados desses algoritmos, destaca-se a pesquisa de Hicks (2019) sobre o desenvolvimento histórico de sistemas públicos automatizados no Reino Unido e a introjeção da discriminação contra pessoas transgênero, de forma frequentemente intencional e sob o manto de maior eficiência:

> A fim de aumentar sua própria eficiência e poder, tais sistemas devem representar a realidade e traduzi-la em um cenário informativo onde se possa atuar sobre ela de forma aparentemente sem atritos, desinteressada e sem vieses. Na verdade, porém, esse processo de tornar a informação computável depende da institucionalização das visões e dos preconceitos daqueles que constroem o sistema e, reflexivamente, serve aos seus fins. (HICKS, 2019, p.29-30, tradução nossa).

Passando então à investigação sobre as consequências apontadas na literatura para o fenômeno do viés algorítmico, há um achado interessante: em geral essas consequências não são especificadas, ficando implícito que viés e discriminação são os próprios efeitos deletérios, injustos ou violadores de direitos humanos (KAMISHIMA *et al.*, 2018; MCGREGOR; MURRAY; NG, 2019; OLHEDE; WOLFE, 2018; KIRKPATRICK, 2016; NOBLE, 2018). Barocas e Selbst (2016, p.674, tradução nossa), além de mencionarem desigualdades de acesso e de participação na sociedade em virtude da reprodução de vieses sociais, sintetizam tal tendência: "Abordada levianamente, a mineração de dados pode reproduzir padrões de discriminação existentes, herdar o preconceito dos tomadores prévios de decisão, ou simplesmente refletir os preconceitos generalizados que persistem na sociedade".

Alguns autores abordam o papel do viés algorítmico no aprofundamento de desigualdades pré-existentes (ECKHOUSE *et al.*, 2019; GUTA; VORONKA;

[17] Sobre essa questão ver também Mayson (2018).

GAGNON, 2018; REGAN; JESSE, 2019; SEVIGNANI, 2017), por vezes destacando a retroalimentação da discriminação por vias algorítmicas: "O viés social e o viés algorítmico podem reforçar-se mutuamente em um *feedback loop* – um círculo vicioso de injustiça acelerado por nossas ferramentas de *big data*" (AYRE; CRANER, 2018, p.344, tradução nossa)[18]. Outros estudos ressaltam os perigos intrínsecos de discriminação em quaisquer categorizações e perfilamentos (ECKHOUSE *et al.*, 2019; RICHARDS, 2012; LEPRI *et al.*, 2018).

Potenciais soluções para o viés algorítmico são investigadas na maior parte de todos os artigos aqui avaliados: somente 10 artigos não as mencionam. Em especial, as soluções exploradas focam em temas de natureza normativa (i.e., em regras de conduta centradas nos criadores e/ou usuários das ferramentas) – em transparência (às vezes associada à auditabilidade), em privacidade, em possibilidade de monitoramento, em códigos de boas condutas e responsabilização (*accountability*) (KIRKPATRICK, 2016; YU; ALÌ, 2019; COTINO HUESO, 2019; RICH, 2016; STRANDBURG, 2019). Além disso, alguns estudos discutem técnicas para modelagem de algoritmos, como o uso de paradigmas específicos, garantia de *inputs* diversos, documentação dos passos tomados, dentre outros (EDWARDS; RODRIGUEZ, 2019; HAZAN, 2013; STRAUß, 2018; HILDEBRANDT; KOOPS, 2010). Em outros casos, soluções sociais foram analisadas, como maior envolvimento comunitário no desenvolvimento das ferramentas e o estímulo a times multidisciplinares (HOWARD; BORENSTEIN, 2018; PASQUALE, 2017; CRAMER *et al.*, 2018).

Há ponderação por parte de alguns autores quanto à inevitabilidade de vieses nos algoritmos. Nestes estudos se propõe sua mitigação ao invés de tentar solucionar o problema, além da proposição de soluções não necessariamente algoritmocêntricas (AYRE; CRANER, 2018; MBADIWE, 2018).

Mayson (2018) e Cotino Hueso (2019) propõem o uso de dados de forma a evitar positiva e ativamente vieses, configurando uma inversão do fenômeno ausente no resto da literatura: "Empregadores e pesquisadores podem usar dados para diagnosticar onde e como os vieses cognitivos ou estruturais estão atualmente operando de forma prejudicial aos grupos desfavorecidos" (KIM, 2017, p.865, tradução nossa). Os dados, então, seriam essenciais para projetar situações sem vieses.

Por fim, é patente a importância dada à transparência. De todos os temas normativos que foram quantificados, este foi o mais mencionado, presente em 53 artigos[19]. Grande parte da literatura a defende, ainda que com ressalvas e múltiplas

[18] Sobre essa questão ver também Mann; Matzner (2019).

[19] Além de transparência, dentre os outros temas normativos quantificados, justeza foi mencionada em 45 artigos, responsabilização em 35 artigos, responsabilidade (*liability*) em 35, e privacidade em 34 artigos.

considerações quanto à abordagem prática. A transparência parece ser vista como uma estratégia para permitir a auditabilidade e o monitoramento de algoritmos, facilitando a detecção e prevenção de vieses:

> Expandir o foco do debate sobre inteligibilidade [*explainability*] para incluir a responsabilização pública é, portanto, apenas o primeiro passo para uma visão mais realista sobre as diversas ramificações da questão da inescrutabilidade das ferramentas decisórias. Antes de incorporar essas ferramentas decisórias baseadas em aprendizagem de máquina […] os tomadores de decisão devem ter uma visão clara sobre quais informações estão disponíveis para todos os atores do sistema. Isso lhes permitiria avaliar se essa informação, combinada com outros mecanismos, poderia fornecer responsabilização e coordenação suficientes para justificar o uso de uma determinada ferramenta de decisão automatizada em um determinado contexto (STRANDBURG, 2019, p.1857, tradução nossa).

Os críticos à transparência questionam sua viabilidade em vista de promover desestímulos econômicos (ATKINSON, 2018) ou por sua insuficiência para fazer frente aos vieses, os quais muitas vezes são oriundos de estruturas sociais dissociadas da obtenção de informações claras sobre o algoritmo (MBADIWE, 2018). Em duas pesquisas empíricas, autores argumentam que o viés algorítmico detectado não teria sido resolvido com maior transparência (LAMBRECHT; TUCKER, 2019; THELWALL, 2018).

É relevante mencionar que o enviesamento algorítmico não foi o principal tema tratado por 84 artigos (68,3%), e que parte significativa dos estudos discute o viés algorítmico no contexto de outro tema, como, por exemplo: a) análise de estatutos jurídicos (geralmente dos EUA) em relação a diferentes usos de algoritmos; b) viés de gênero na visualização de anúncios na internet (LAMBRECHT; TUCKER, 2019); c) o ranqueamento de resultados do *Google Search* (GILLESPIE, 2017); d) mecanismos de moderação e incentivos em vídeos do Youtube (BISHOP, 2018); e) os escores de crédito e saúde (PASQUALE, 2019); f) tecnologias "vestíveis" (*wearables*) (MONTGOMERY; CHESTER; KOPP, 2018, SAVIRIMUTHU, 2017); g) carros autônomos (LIM; TAEIHAGH, 2019).

Em suma, o típico texto sobre viés algorítmico é um ensaio recente, publicado em periódico jurídico e oriundo de acadêmicos vinculados a universidades norte-americanas ou europeias. Amiúde, tais ensaios buscam introduzir o tema por meio de exemplos, analisá-los em conjunto com determinado contexto, e explorar perfunctoriamente seus efeitos sociais e jurídicos, assim como possíveis soluções (em especial a transparência).

Tal predominância pode derivar do caráter recente do tema, aliado à preocupação crescente com seus efeitos discriminatórios. Foi recorrente a menção ao COMPAS, *software* utilizado para notas de risco em réus de processos penais nos EUA. Em 2016, a Organização Não-Governamental ProPublica publicou uma análise extensiva, hoje célebre, dos resultados desse *software*, apontando vieses raciais em seus resultados (ANGWIN; LARSON; MATTU; KIRCHNER, 2016). É possível que o caso tenha contribuído para maior conscientização e estímulo para pesquisas sobre o tema. Esta hipótese encontra respaldo no fato de que viés racial (*racial bias*) é mencionado em 38 artigos (30,9%), seguido de viés de gênero (*gender bias*) em 19 (15,4%) dos artigos aqui avaliados.

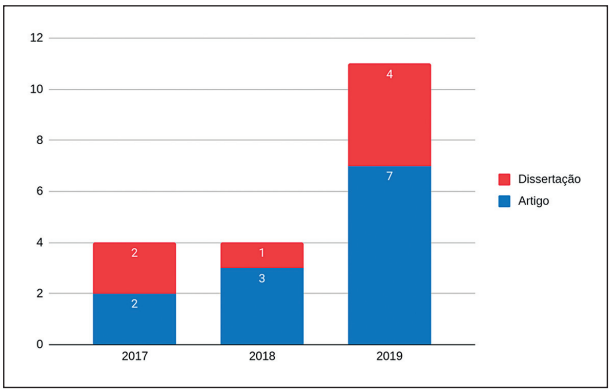
## IV. Literatura Nacional

Conforme mencionado na seção de metodologia, foram raras as entradas de artigos sobre viés algorítmico em português. Miskolci e Balieiro (2018) notam que, apesar de o termo Sociologia Digital ter sido usado apenas recentemente no Brasil (por ocasião de dossiê específico publicado em 2016), tal discussão não é nova. Eles também argumentam que se tem aprofundado a reflexão sobre o uso de tecnologias de comunicação em rede. Ao apontarem seis grandes eixos temáticos da área[20], elencam enquanto temática necessária a "[recuperação da] perspectiva da moldagem social da tecnologia, a qual reconhece sua historicidade e o caráter criativo e aberto de seus usos, usos que variam socialmente e só podem ser aferidos por meio de investigação e análise de como eles se inserem nas práticas cotidianas" (MISKOLCI; BALIEIRO, 2018, p.150).

É justamente esta perspectiva que permite a discussão do viés algorítmico. A trajetória nos estudos de tecnologias digitais e sociedade nas Ciências Sociais brasileiras chama atenção para o caráter recente desses estudos e ajuda a dar sentido às escassas produções encontradas durante o levantamento aqui realizado. Todas as entradas em português se deram nos últimos três anos (Gráfico 6).

[20] Seriam eles ciberativismo, mídias digitais e afetividades, ciberespaço, práticas sociais digitais, desigualdade digital e reflexões metodológicas (MISKOLCI; BALIEIRO, 2018, p.138-139).

**Gráfico 6 –** Ano de publicação das entradas nacionais

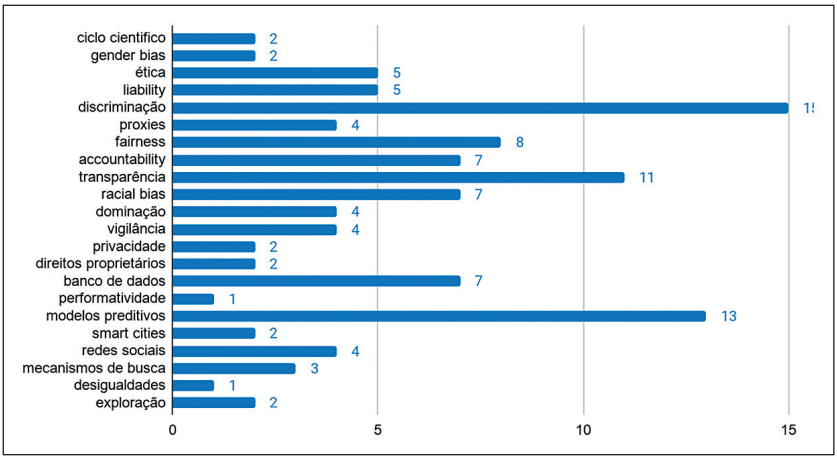



Fonte: Elaboração própria.

A exemplo da desigualdade geopolítica na produção de artigos que tratam de viés algorítmico, a produção nacional não escapa das desigualdades regionais, observadas também na revisão de Sampaio *et al.* (2018) sobre a procedência dos artigos sobre internet e política. Aqui, os trabalhos se concentram no Sudeste (15 entradas), Sul (3) e Centro-Oeste (2), sem entradas das regiões Norte e Nordeste.

Em consonância com a produção internacional, os trabalhos brasileiros também convergiram em três grandes áreas: Direito, Estudos Interdisciplinares e Comunicação. As questões trabalhadas nos artigos têm forte influência do campo jurídico e de governança, com alguma discussão técnica:

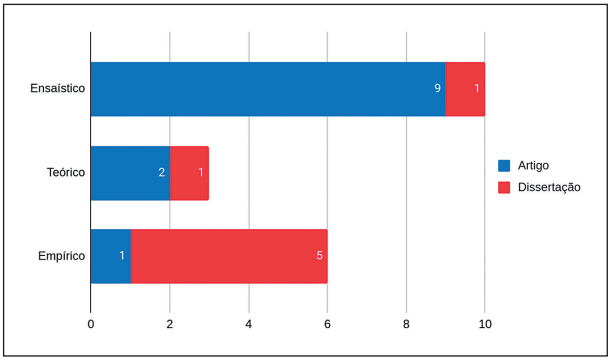
**Gráfico 7 –** Palavras-chave das entradas nacionais




Fonte: Elaboração própria.

Em termos agregados, predominam os trabalhos de caráter ensaístico. Contudo, viu-se que esta distribuição não é uniforme:

**Gráfico 8 –** Tipo textual das entradas nacionais




Fonte: Elaboração própria.

É de destacar um esforço de exposição teórica e bibliográfica de temas relacionados à inserção dos algoritmos na vida social. Marcadamente, os trabalhos se esmeram em descrever o processo de elaboração e funcionamento dos algoritmos, inteligência artificial, aprendizado de máquina e *big data*. Ainda que não de forma explícita, aparecem questões vinculadas ao status ontológico dos algoritmos, à distribuição de responsabilidade e à agência dessas tecnologias.

No que tange às principais questões levantadas, alguns trabalhos abordam a problemática da disseminação de algoritmos nas tomadas de decisão no contexto brasileiro, a partir de experiências de outras localidades, ou de discursos transnacionais (SILVA, 2019; SILVEIRA; MOURA; ALMEIDA, 2019; ARAÚJO, 2017; LUCENA, 2019; BRANCO, 2019); em geral, esses estudos se baseiam em dados secundários. Outros, de caráter empírico, têm os olhos voltados exclusivamente para experiências nacionais (SAMPAIO, 2017; WERNER, 2019; MOURA, 2018).

Recortando a amostra para os artigos cujo foco é o viés algorítmico, restam oito dos 12 artigos analisados. Acompanhando a tendência internacional de indefinição do termo, (FAVARETTO; DE CLERCQ; ELGER, 2019), apenas quatro desses apresentam alguma caracterização do fenômeno, dividindo-se entre uma visão técnica do viés algorítmico – que o caracteriza como um processo de generalização a partir de uma categoria que gera consequências consideradas socialmente discriminatórias (LINDOSO, 2019; FERRARI; BECKER; WOLKART, 2018) – e uma visão normativa – que o equaliza com o tratamento discriminatório (LUCENA, 2019; SILVEIRA; MOURA; ALMEIDA, 2019). São notáveis os esforços de distinção de termos como discriminação algorítmica, perfilamento e viés

algorítmico feitos por Mattiuzzo (2019) e Mendes e Mattiuzzo (2019), na medida em que, embora sejam todos processos de categorização e generalização, uns são moralmente condenáveis (e assim, discriminatórios), outros não.

Na discussão sobre as causas de viés algorítmico, destaca-se menções à estrutura social de onde derivam os dados, o enviesamento dos dados de amostra e os diversos passos de desenvolvimento da ferramenta. Apenas um artigo reflete sobre o viés decorrente da implementação e aplicação dessas ferramentas.

Nos textos nacionais, a abordagem das consequências do uso de sistemas algorítmicos enviesados privilegia o problema da discriminação, bem como a formalização de estereótipos e a hipervigilância sobre determinados grupos. Ademais, alguns autores mencionam os riscos de incorrer em *feedback loops*, de aumento da desigualdade, de violação de direitos e de descentralização do elemento humano do processo decisório.

Como medidas para a mitigação do viés algorítmico são apontadas disposições de cunho moral, normativo, pedagógico, técnico e social. As medidas mais mencionadas são a adoção de princípios éticos no desenvolvimento desses sistemas e o estabelecimento de mecanismos de auditoria, seguidas da sugestão de regulação normativa apropriada e da composição multidisciplinar do time de desenvolvedores. Outras providências elencadas são governança, criação de órgãos públicos de fiscalização, manutenção de *proxies* positivos e fomento de iniciativas para desconstrução das pretensões de objetividade que circundam sistemas algorítmicos.

Em suma, a literatura nacional se caracteriza por trazer os termos do debate internacional e refletir sobre sua aplicabilidade no contexto brasileiro, potencialidades e limites, com rara proposição de novas questões, perspectivas teóricas ou inovações metodológicas. Até o momento, são poucos os estudos empíricos brasileiros sobre viés algorítmico e, como notam Miskolci e Balieiro (2018), o questionamento do determinismo tecnológico ainda não é tema resolvido.

A adoção de tecnologias digitais na esfera pública vem ocorrendo de forma desorganizada e difusa (ARBIX; MIRANDA, 2015). Ainda assim e talvez por isso mesmo, o fenômeno do viés algorítmico começa se evidenciar no Brasil, ao mesmo tempo que desperta gradativamente os interesses públicos (para sua implantação) e acadêmicos (para a compreensão de seus mecanismos, causas, consequências e limitações). O próprio estudo sobre os processos de implementação dessas tecnologias tem o potencial de agregar tanto teórica como metodologicamente.

## V. Discussão

A análise do *corpus* de trabalhos selecionados desperta questões interessantes. Na literatura internacional, predominam os trabalhos do Norte Global, em especial

dos EUA e Europa, de produção concentrada nos últimos quatro anos. São obras de natureza recorrentemente ensaística, com ainda baixo desenvolvimento teórico e empírico sobre o tema. Ao mesmo tempo, o viés algorítmico aparece como tópico secundário em artigos dedicados a outros objetos.

O caráter ensaístico e a novidade do debate nas Humanidades podem explicar, em parte, a constatação da ambiguidade do termo, assim como sua aproximação com outras expressões, como discriminação algorítmica (ou ainda, embora menos frequente, de perfilamento e filtragem algorítmica). Ora tratados como sinônimos, ora como relação causal, discriminação aparece como a principal palavra-chave no *corpus* de textos.

A isso pode ser agregada outra questão, que é a recorrência da construção do modelo como a causa mais citada do viés algorítmico. Conforme já discutido, seja pela literatura abordada neste artigo, seja nos estudos sobre ciência e tecnologia de maneira mais ampla, a tecnologia é um produto social, na medida em que a sociedade interfere nas suas condições de produção e circulação, mas também na constituição dos valores e subjetividades dos agentes que as produzem.

Por vezes, o viés algorítmico aparece nos textos como viés subjetivo, em contraponto à pretensa objetividade da ferramenta. O foco nos processos pelos quais a subjetividade se transfere e contamina a ferramenta parece encontrar eco também no tratamento e denominação normativa que circunda o viés algorítmico, isto é, a discriminação. De um lado, há uma dimensão moral e ética que está envolvida nas formas de categorização e que vai destacar a discriminação como forma inaceitável de outros processos de categorização e filtragem. De outro, há no uso de discriminação pela perspectiva normativa, a busca pela intencionalidade do sujeito e pela ação discriminatória, uso este de legado individualista e que parte de uma concepção liberal de ação[21].

Os debates sobre discriminação nos estudos sobre relações raciais, por exemplo, destacam o giro interpretativo, que passou da análise centrada na conduta individual e intencional (uma pessoa que discrimina outra, intencionalmente), para a análise centrada nos mecanismos sociais pelos quais a desigualdade é reproduzida, ainda que de maneira não-intencional (processos que levam a resultados discriminatórios). Ao que tudo indica, pensar o viés algorítmico a partir da discriminação significa pensar a tecnologia a partir do agente que a constrói.

Por um lado, este parece ser um movimento que responde à construção da mitologia envolvendo o uso de *big data*[22] e ferramentas de automatização como

[21] Inclusive, esta é reconhecidamente uma dificuldade no campo jurídico, ao buscar configurar uma discriminação intencional e, com isso, possibilitar a aplicação de leis anti-discriminatórias sem incorrer no tratamento não-igualitário (BAROCAS; SELBST, 2016; para o caso brasileiro, ver MACHADO; LIMA; SANTOS, 2019).

[22] Ver nota de rodapé 9.

tomadores de decisão mais objetivos. Por outro, pode descuidar de uma compreensão mais abrangente sobre as posições dos agentes. Para além da análise e desconstrução de discursos que ganham amplo espaço no debate público, a investigação científica deve se colocar também a tarefa de propor novas formas de compreensão dos fenômenos que estuda. Nesse sentido, a Sociologia enquanto disciplina tem muito a contribuir como ferramenta de interrogação do fenômeno do viés algorítmico.

## VI. Considerações finais

Este artigo produziu um inventário provisório da ocorrência de estudos acerca do viés algorítmico, objeto cujo estudo se encontra em expansão. A escolha pela realização de um levantamento bibliográfico com caducidade precoce justifica-se pelo interesse na disseminação das principais reflexões em língua portuguesa sobre um tema de crescente relevância e interesse para a realidade social brasileira.

Após apresentar algumas tendências identificadas neste *corpus* de trabalhos, chama-se a atenção para algumas questões de interesse sociológico, que consistem em como pensar tecnologias digitais a partir de uma perspectiva compreensiva, que busque integrar análises micro e macroprocessuais.

Origens possíveis do viés algorítmico são atribuídas a questões presentes em todas as etapas do desenvolvimento de um sistema algorítmico. Dentre essas, é recorrente a busca pela subjetividade embutida no sistema: a subjetividade do desenvolvedor – que escolhe e trata os bancos de dados, que confere peso e testa os modelos –, a subjetividade do implementador ou daquele que adota o sistema no seu processo decisório. A maioria dessas entradas, inclusive por estarem tão estreitamente ligadas às noções jurídicas de discriminação, centra-se no agente. Enquanto plano reflexivo, não menos importante é o reconhecimento dos elementos que transcendem a subjetividade dos atores envolvidos. Pensar na discriminação e concentrar-se na busca pelo autor do viés pode significar a negligência da desigualdade estrutural presente nos dados, ou do contexto no qual esses indivíduos, desenvolvedores, estão posicionados.

Mesmo na hipotética situação de acessarmos um banco de dados representativo, retornamos à questão das relações de desigualdades fundantes de nosso mundo social, com a qual historicamente nos debatemos. Para além da correção do viés subjetivo, retornamos a um debate existente, ainda que pouco explorado neste *corpus* de trabalhos: a reprodução das desigualdades sociais – questões essas que embasam, por exemplo, as literaturas sobre categorização, estratificação e desigualdades sociais e raciais. Lacuna essa que poderá ser rapidamente preenchida com o adensamento do debate sobre o tema e com novas contribuições, teóricas e empíricas.

## *Algorithmic bias: a provisional review*

***ABSTRACT:*** *This paper stands in the field of Digital Sociology and proposes to carry out a provisional bibliographic review on algorithmic bias, based on the main journals databases in English, Spanish and Portuguese. Our aim is to assess how the debate on algorithmic bias in the Humanities has developed, observing which definitions, origins, diagnostics and perspectives are presented for the phenomenon. It is observed that the majority of articles are of essayistic nature and produced in the Global North, with a low penetration of the subject in Spanish and Portuguese language literature. The expression remains broadly undefined, and is sometimes treated as synonymous with discrimination, sometimes as its cause. The main sources of bias are, according to the literature, the model development and its training data, which largely lead to recommendations for increasing transparency about the process and suggest an analytical tendency to emphasize the subjective character of algorithmic bias. There seems to be an analytical tendency in the subjective character of bias. As a perspective, we point out the importance of integrating the elements transcending the subjectivity of these actors to this analysis.*

***KEYWORDS:*** *Algorithmic bias. Algorithm. Discrimination. Digital Sociology.*

## *Sesgo algorítmico: una síntesis provisional*

***RESUMEN:*** *Este artículo se sitúa en el campo de la Sociología Digital y propone realizar una revisión bibliográfica provisional sobre el sesgo algorítmico, basada en los principales portales de revistas en inglés, español y portugués. El objetivo es seguir el desarrollo del debate sobre el sesgo algorítmico en las Humanidades, observando las definiciones, los orígenes, los diagnósticos y las perspectivas presentadas para el fenómeno. El resultado fue que la mayoría de los artículos es de naturaleza ensayística, producidos en el Norte Global, con baja penetración del tema en la literatura en español y portugués. Hay una gran falta de definición del término, que por veces se trata como sinónimo de discriminación, por veces como su causa. Las principales fuentes de sesgo presentadas son la construcción de la herramienta y sus datos de capacitación, lo que en gran medida da lugar a sugerencias para aumentar la transparencia del proceso y sugiere una tendencia analítica a subrayar el carácter subjetivo del sesgo algorítmico. Parece haber una tendencia analítica en el carácter subjetivo del sesgo. Como perspectiva, señalamos la importancia de integrar en el análisis elementos que trasciendan la subjetividad de estos actores.*

***PALABRA CLAVE:*** *Sesgo algorítmico. Algoritmo. Discriminación. Sociología digital.*

## Agradecimentos

Os autores agradecem ao prof. Leopoldo Waizbort pelos comentários, às pesquisadoras do Núcleo de Estudos da Violência Débora Piccirillo e Roberta Novello, à Mônica Corso e à colega Jéssica Höring pela leitura; por fim, ao Grupo de Pesquisa em Sociologia Digital da FFLCH/USP. Eventuais incorreções são de inteira responsabilidade dos autores. Somos gratos pelo financiamento da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processo nº 2019/02612-0.

## REFERÊNCIAS

ALPAYDIN, Ethem. **Machine learning**: the new AI. Cambridge, MA: MIT press, 2016.

ANGWIN, Julia; LARSON, Jeff; MATTU Surya; KIRCHNER Lauren. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. **ProPublica**. May 23, 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 25 jun. 2020.

ARAÚJO, Camila. Identifying stereotypes in the online perception of physical attractiveness. Dissertação (Mestrado em Ciência da Computação). Orientador: Wagner Meira Júnior. 82f. Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, 2017.

ARBIX, Glauco; MIRANDA, Zil. Inovação em tempos difíceis. **Plural**, v.22, n.2, p.18-36, 2015.

ATKINSON, Robert. "It Is Going to Kill Us!" and Other Myths About the Future of Artificial Intelligence. **IUP Journal of Computer Sciences**, v.12, n.4, p.7-56, 2018.

AYRE, Lori; CRANER, Jim. Algorithms: avoiding the implementation of institutional biases. **Technology Column**. v.37, n.3, p.341-347, 2018.

BAROCAS, Solon; SELBST, Andrew D. Big data's disparate impact. **Calif. L. Rev.**, v.104, p.671-732, 2016.

BENJAMIN, Ruha. **Captivating Technology**: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life. Durham: Duke University Press, 2019a.

BENJAMIN, Ruha. **Race After Technology**: Abolitionist Tools for the New Jim Code. Cambridge: Polity, 2019b.

BENJAMIN, Ruha. Innovating inequity: if race is a technology, postracialism is the genius bar. **Ethnic and Racial Studies**, v.39, n.13, p.2227-2234, 2016.

BISHOP, Sophie. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. **Convergence**, v.24, n.1, p.69-84, 2018.

BOYD, Danah; CRAWFORD, Kate. Critical questions for Big Data. **Information**, **Communication**, **Society**. York, v.15, n.5, p.662-679, 2012.

BRANCO, Priscilla. Smart Cities como dispositivos biopolíticos. **VI Simpósio Internacional Lavits**, Salvador, 2019.

BURK, Dan L. Algorithmic Fair Use. **U. Chi. L. Rev.**, v.86, p.283-308, 2019.

CASTRO, Julio Cesar. Redes sociais como modelo de governança algorítmica. **Matrizes**. v. 12, p.165-191, 2018.

COTINO HUESO, Lorenzo. Derecho y garantías ante el uso público y privado de inteligencia artificial, robótica y big data. *In*: BAUZÁ, Marcelo (org.). **El Derecho de las TIC en Iberoamérica**. Montevideo: La Ley- Thompson-Reuters, p.917-952, 2019.

CRAMER, Henriette; GARCIA-GATHRIGHT, Jean; SPRINGER, Aaron; REDDY, Sravana. Assessing and addressing algorithmic bias in practice. **Interactions**, v.25, n.6, p.58-63, 2018.

CUSTERS, Bart; CALDERS, Toon; SCHERMER, Bart.; ZARSKY, Tal (Eds.). **Discrimination and Privacy in the Information Society**: Data Mining and Profiling in Large Databases. Springer, 2012.

DANKS, David; LONDON, Alex John. Algorithmic Bias in Autonomous Systems. **IJCAI** 17, p.4691-4697, 2017.

DESROSIÈRES, Alain. **The politics of large numbers**: a history of statistical reasoning. Cambridge, Mass.: Harvard University Press, 2011.

DIAKOPOULOS, Nicholas. Accountability in Algorithmic Decision Making. **Comm**. **ACM,** 2016.

DIXON-ROMÁN, Ezekiel; NICHOLS, Philip; NYAME-MENSAH, Ama. The racializing forces of/in AI educational. **Technologies**, **Learning**, **Media and Technology**: Special Issue on "AI and Education: critical perspectives and alternative futures". p.1-15, 2019.

ECKHOUSE, Laurel; LUM, Kristian; CONTI-COOK, Cynthia; CICCOLINI, Julie. Layers of bias: A unified approach for understanding problems with risk assessment. **Criminal Justice and Behavior**, v.46, n.2, p.185-209, 2019.

EDWARDS, John Steven; RODRIGUEZ, Eduardo. Remedies against bias in analytics systems. **Journal of Business Analytics**, v.2, n.1, p.74-87, 2019.

EUBANKS, Virginia. **Automating inequality**: how high-tech tools profile, police, and punish the poor. New York: St. Martin's Press, 2018.

FAVARETTO, Maddalena; DE CLERCQ, Eva; ELGER, Bernice. Big Data and discrimination: perils, promises and solutions. A systematic review**. Journal of Big Data**, v.6, n.1, p.1-27, 2019.

FERRARI, Isabela; BECKER, Daniel; WOLKART, Erik. Arbitrium ex machina: panorama, riscos e a necessidade de regulação das decisões informadas por algoritmos. **Revista dos Tribunais online**. v.995, p.1-16, 2018. Disponível em: https://www.academia.edu/38199022/ARBITRIUM_EX_MACHINA_PANORAMA_RISCOS_E_A_NECESSIDADE.pdf. Acesso em: 26 jun. 2020.

FOURCADE, Marion; HEALY, Kieran. Seeing like a Market. **Socio-Economic Review**, v.15, n.1, p.9-19, 2016.

FRIEDMAN, Batya; NISSENBAUM, Helen. Bias in computer systems. **ACM Transactions on Information Systems**. v.14, n.3, p.330-347, 1996.

GILLESPIE, Tarleton. A relevância dos algoritmos. **Revista Parágrafo**, v.6, n.1, p.95-121, 2018.

GILLESPIE, Tarleton. Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. **Information**, **Communication**, **Society**, v.20, n.1, p.63-80, 2017.

GRAHAM, Stephen; WOOD, David. Digitizing Surveillance: Categorization, Space, Inequality. **Critical Social Policy**. v.23, n.2, p.227-248, 2003.

GUTA, Adrian; VORONKA, Jijian; GAGNON, Marilou. Resisting the digital medicine Panopticon: toward a bioethics of the oppressed. **The American Journal of Bioethics**, v.18, n.9, p.62-64, 2018.

HACKING, Ian. **The taming of chance**. New York: Cambridge University Press, 1990.

HAZAN, Joshua G. Stop being evil: A proposal for unbiased Google search. **Mich. L. Rev.**, v.111, p.789, 2013.

HICKS, Marie. Hacking the Cis-tem. **IEEE Annals of the History of Computing**, v.41, n.1, p.20-33, 2019.

HILDEBRANDT, Mireille; KOOPS, Bert-Jaap. The Challenges of Ambient Law and Legal Protection in the Profiling Era. **The Modern LR**, v.73, n.2, p.428-460, 2010.

HUQ, Aziz Z. Racial equity in algorithmic criminal justice. **Duke LJ**, v.68, p.1043-1134, 2019.

HOWARD, Ayanna; BORENSTEIN, Jason. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. **Science and Engineering Ethics**, v.24, n.5, p.1521-1536, 2018.

KAMISHIMA, Toshihiro; SHOTARO, Akaho; ASOH, Hideki; SAKUMA Jun. Model-based and actual independence for fairness-aware classification. **Data mining and knowledge discovery**, v.32, n.1, p.258-286, 2018.

KIM, Pauline. Data-Driven Discrimination at Work. **William and Mary LR**. v.58, n.3, p. 856-936, 2017.

KIRKPATRICK, Keith. Battling algorithmic bias: how do we ensure algorithms treat us fairly? **Comm. of the ACM**, v.59, p.16-17, 2016.

LA FORS, Karolina; CUSTERS, Bart; KEYMOLEN, Esther. Reassessing values for emerging big data technologies: Integrating design-based and application-based approaches. **Ethics and Information Technology**, v.21, p.209-226, 2019.

LAMBRECHT, Anja; TUCKER, Catherine. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. **Management Science**. v.65, n.7, p.2966-2981, 2019.

LE GALÈS, Patrick. Performance measurement as a policy instrument. **Policy Studies**. v.37, n.6, p.508-520, 2016.

LEPRI, Bruno; NURIA, Oliver; LETOUZÉ, Emmanuel; PENTLAND, Alex; VINCK, Patrick. Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. **Philosophy**; **Technology**. v.31, n.10, p.611-627, 2018.

LIM, Hazel; TAEIHAGH, Araz. Algorithmic Decision-Making in AVs: Understanding Ethical and Technical Concerns for Smart Cities. **Sustainability**. v.11, n.20, 2019.

LINDOSO, Maria C. Discriminação de gênero em processos decisórios automatizados. Orientadora: Ana Frazão. 116f. Dissertação (Mestrado em Direito). Faculdade de Direito, Universidade Nacional de Brasília, 2019.

LUCENA, Pedro A. Policiamento preditivo, discriminação algorítmica e racismo: potencialidades e reflexos no Brasil. **VI Simpósio Internacional Lavits**, Salvador, 2019.

LUPTON, Deborah. **Digital sociology**. New York: Routledge, 2015.

McGREGOR, Lorna; MURRAY, Daragh; NG, Vivian. International Human Rights Law as a framework for algorithmic accountability. **International and Comparative Law Quarterly**. v.68, n.2, p.309-343, 2019.

MACHADO, Henrique F. Algoritmos, regulação e governança: uma revisão de literatura. **Revista de Direito Setorial e Regulatório**. Brasília: v.4, n.1, p.39-62, 2018.

MACHADO, Marta; LIMA, Márcia R; SANTOS, Natália. Anti-racism legislation in Brazil: the role of the Courts in the reproduction of the myth of racial democracy. **Revista de Investigações Constitucionais**, Curitiba, vol.6, n.2, p.267-296, 2019.

MANN, Monique; MATZNER, Tobias. Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. **Big Data**; **Society**, v.6, n.2, n.p., 2019.

MARRES, Noortje. **Digital Sociology**: The Reinvention of Social Research. Oxford: Polity Press, 2017.

MATTIUZZO, Marcela. Algorithms and Big Data: Considerations on Algorithmic Governance and Its Consequences for Antitrust Analysis. **Revista de Economia Contemporânea**. Rio de Janeiro v.23, n.2, p.1-19, 2019.

MAYSON, Sandra G. Bias in, bias out. **Yale LJ**, v.128, p.2218-2300, 2018.

MBADIWE, Tafari. Algorithmic injustice. **The New Atlantis**, n.54, p.3-28, 2018.

MENDES, Laura; MATTIUZZO, Marcela. Discriminação algorítmica: conceito, fundamento legal e tipologia. **Direito Público**, v.16, n.90, p.39-64, 2019.

MISKOLCI, Richard; BALIEIRO, Fernando. Sociologia Digital: balanço provisório e desafios. **Revista Brasileira de Sociologia**. Belo Horizonte: v.6, n.12, p.132-156, 2018.

MONTGOMERY, Kathryn; CHESTER, Jeff; KOPP, Katharina. Health wearables: ensuring fairness, preventing discrimination, and promoting equity in an emerging Internet-of-Things environment. **Journal of Information Policy**, v.8, p.34-77, 2018.

MOURA, Carolina. Associações sociotécnicas: mediações algorítmicas e a economia das ações no Facebook. Orientadora: Suely Henrique de Aquino Gomes. 162f. Dissertação (Mestrado em Comunicação). Faculdade de Informação e Comunicação, Universidade Federal de Goiás, 2018.

NOBLE, Safiya. **Algorithms of Oppression: How Search Engines Reinforce Racism.** New York: New York University Press, 2018.

OLHEDE, S. C.; WOLFE, P. J. The growing ubiquity of algorithms in society: implications, impacts and innovations. **Philosophical Transactions of the Royal Society A**: Mathematical, Physical and Engineering Sciences, v.376, n.2128, p.1-16, 2018. Disponível em: http://dx.doi.org/10.1098/rsta.2017.0364. Acesso em: 25 jun. 2020.

O'NEIL, Cathy. **Weapons of math destruction**: how big data increases inequality and threatens democracy. New York: Crown, 2016.

PASQUALE, Frank. Data-Informed Duties in AI Development (2019). **Columbia LR**, v.119, p.1917-1940, 2019.

PASQUALE, Frank. A esfera pública automatizada. **LÍBERO**. São Paulo, v.39, p.16-35, 2017.

PASQUALE, Frank. **The Black Box Society**: The Secret Algorithms That Control Money and Information. Cambridge (MA): Harvard University Press, 2015.

POWER, Michael. Counting, Control and Calculation: Reflections on Measuring and Management. **Human Relations**, v.57, n.6, p.765-783, 2004.

REGAN, Priscilla M.; JESSE, Jolene. Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. **Ethics and Information Technology**, v.21, n.3, p.167-179, 2019.

RICH, Michael. Machine learning, automated suspicion algorithms, and the fourth amendment. **University of Pennsylvania Law Review**, v.164, p.871-929, 2016.

RICHARDS, Neil. The dangers of surveillance. **Harv. L. Rev.**, v.126, p.1934-1965, 2012.

SAMPAIO, Alice. Data Brokers: um novo modelo de negócios baseado em vigilância de dados. Orientadora: Marta Kanashiro. 135f. Dissertação (Mestrado em Divulgação Científica e Cultural). Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, 2017.

SAMPAIO, Rafael Cardoso; MITOZO, Isabele; MASSUCHIN, Michele Goulart; FONTES, Giulia Sbaraini; PENTEADO, Cláudio Luis de Camargo. Ciberpolítica, ciberativismo e cibercultura: uma análise dos papers apresentados no grupo de trabalho da Anpocs. **BIB**. v.1, n.85, p.126-147, 2018.

SAVIRIMUTHU, Joseph. Do algorithms dream of 'data' without bodies? **International Review of Law**, **Computers**, **Technology**, v.31, n.2, p.243-262, 2017.

SEVIGNANI, Sebastian. Surveillance, classification, and social inequality in informational capitalism: The relevance of exploitation in the context of markets in information. **Historical Social Research**, v.42, n.1, p.77-102, 2017.

SHIN, Donghee Don; FOTIADIS, Anestis; YU, Hongsik. Prospectus and limitations of algorithmic governance: an ecological evaluation of algorithmic trends. **Digital Policy**, **Regulation and Governance**, 2019.

SILVA, Tarcizio. Visão Computacional e Vieses Racializados: Branquitude Como Padrão No Aprendizado De Máquina. **II COPENE NORDESTE**. Anais do Encontro, p.1-13, 2019.

SILVEIRA, Sérgio Amadeu; MOURA, Lucas; ALMEIDA, Lucas Theodoro. A reprogramação da sociedade nos discursos sobre algoritmos. **VI Simpósio Internacional Lavits**, Salvador, 2019.

STRANDBURG, Katherine. Rulemaking and Inscrutable automated decision tools. **Columbia LR**. v.119, n 7, p.1851-1886, 2019.

STRAUß, Stefan. From Big Data to Deep Learning: A Leap Towards Strong AI or 'Intelligentia Obscura'? **Big Data and Cognitive Computing**, v.2, n.3, p.1-19, 2018.

THELWALL, Mike. Gender bias in sentiment analysis. **Online Information Review**, 42 (1), p.45-57, 2018. Disponível em: http://hdl.handle.net/2436/620633. Acesso em: 26 jun. 2020.

WERNER, Deivid A. A quarta revolução industrial e a inteligência artificial: um estudo sobre seus conceitos, reflexos e possível aplicação no Direito por meio da análise de texto jurídico como forma de contribuição no processo de categorização preditiva de acórdãos. Orientador: Wilson Engelmann. 211f. Dissertação (Mestrado em Direito). UNISINOS, 2019.

YU, Ronald; ALÌ, Gabriele. What's Inside the Black Box? AI Challenges for Lawyers and Researchers. **Legal Information Management**, v.19, n.1, p.2-13, 2019.

ZUBOFF, Shoshana. **The Age of Surveillance Capitalism**: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs, 2019.

ZUBOFF, Shoshana. Big Other: Capitalismo de vigilância e perspectivas para uma civilização da informação. *In*: BRUNO, Fernanda; CARDOSO Bruno; KANASHIRO, Marta; GUILHON, Luciana; MELGAÇO, Lucas (org.). **Tecnopolíticas da vigilância**: Perspectivas da margem. São Paulo: Boitempo, 2018, p.17-68.

Recebido em 05/03/2020.

Aprovado em 31/05/2020.