DOI: https://doi.org/10.26512/rfmc.v8i3.34363

É Possível Evitar Vieses Algorítmicos?*

[Is It Possible to Avoid Algorithmic Bias?]

Carlos Henrique Barth**

Resumo: Técnicas de inteligência artificial (IA) são utilizadas para modelar as atividades humanas e gerar predições comportamentais. Estes sistemas têm apresentado vieses diversos, inclusive de raça e gênero, tipicamente tomados como problemas de engenharia. Realiza-se aqui um esforço argumentativo para mostrar que: 1) escapar dos vieses demanda um sistema que compreenda a estrutura das atividades humanas e; 2) criar um sistema que apresente tal compreensão demanda a solução de problemas fundacionais da IA, em particular, o problema de como modelar o senso comum. No caso de plataformas informacionais que usam desses modelos para intermediar interações com seus usuários, ignorar estes problemas dá margem a uma ilusão de progresso, em que uma crescente influência sobre nosso comportamento é tomada como uma crescente acurácia preditiva. Nesse cenário, argumenta-se que o problema dos vieses está associado a questões não técnicas que devem ser discutidas em espaços públicos.
Palavras-chave: Inteligência Artificial. Vieses Algorítmicos. Governamentalidade Algorítmica.

Abstract: Artificial intelligence (AI) techniques are used to model human activities and predict behavior. Such systems have shown race, gender and other kinds of bias, which are typically understood as technical problems. Here we try to show that: 1) to get rid of such biases, we need a system that can understand the structure of human activities and; 2) to create such a system, we need to solve foundational problems of AI, such as the common-sense problem. Additionally, when informational platforms uses these models to mediate interactions with their users, which is a commonplace nowadays, there is an illusion of progress, for what is an increasingly higher influence over our own behavior is took for an increasingly higher predictive accuracy. Given this, we argue that the bias problem is deeply connected to non-technical issues that must be discussed in public spaces.
Keywords: Artificial Intelligence. Algorithmic Bias. Algorithmic Governance.

*Agradeço à Rochelle Barth, Ernesto Perini, Felipe Nogueira, Eduarda Calado e Samuel Maia por comentários em versões anteriores desse material.

**Mestre em filosofia pela Universidade Federal de Minas Gerais (UFMG). Atualmente realiza doutorado na mesma instituição, com bolsa da CAPES. E-mail: carloshb@protonmail.com. ORCID: https://orcid.org/0000-0002-9327-9818.

I

Com o advento da inteligência artificial (IA), o temor de que a humanidade possa ser dominada por suas próprias criações ganhou novo fôlego. Máquinas de inteligência superior poderiam nos subjugar facilmente, dizem alguns. Porém, já é sabido que o poder de nossos artefatos sobre nós mesmos independe de capacidades sobre-humanas. Somos igualmente vulneráveis a sistemas que não sabem o que fazem. Nesse texto buscamos elucidar, ainda que parcialmente, como isso é possível.

Nos últimos anos, consolidou-se a prática de gerar modelos computacionais das atividades humanas. Seus registros constituem gigantescos volumes de dados (*Big Data*) a partir dos quais é possível, em tese, descobrir novos fatos sobre elas e, portanto, sobre nós. Em particular, tais modelos permitem automatizar a descoberta e a aplicação de categorias. Estas podem ser tão amplas quanto "bom pagador" ou tão específicas quanto "indivíduos que compram sem ponderar muito à noite". Para além de características individuais, claro, esse processo de categorização pode ser aplicado também a comportamentos específicos (adequado, inadequado, suspeito...). Uma vez identificadas, estas categorias permitem realizar predições com um grau de acurácia presumivelmente superior ao que um ser humano seria capaz de alcançar. Qual o risco de conceder liberdade condicional a um detento? Deve um determinado candidato ser contratado? Há chances consideráveis de inadimplência, caso se conceda um empréstimo? Esse é o tipo de pergunta cuja resposta é, cada vez mais, dependente daquilo que os sistemas informacionais têm a dizer.

Não por acaso, há um crescente uso desses modelos para substituir o juízo humano, particularmente em atividades repetitivas.[1] Uma das principais vantagens da automação desses processos é, ou deveria ser, a possibilidade de realizar juízos mais neutros, ou seja, de aplicar os critérios relevantes de modo ponderado e não tendencioso. O que se vê na prática, porém, está longe de ser satisfatório. Não raro, os modelos têm apresentado vieses, atribuindo um peso inadequado a certos critérios de modo sistemático. Os exemplos mais marcantes envolvem vieses de gênero e raça. No caso do COMPAS (*Correctional offender management profiling for alternative sanctions*), ferramenta utilizada nos EUA para categorizar detentos e auxiliar na decisão sobre a concessão de li-

[1] É comum argumentar que o modelo está sendo usado para "auxiliar" o juízo humano, e não para substituí-lo. Raramente esse é o caso, contudo. Se N documentos forem categorizados por um modelo como sendo uma demanda jurídica de tipo X, e se essa categorização não for validada, documento a documento, por um ser humano, o que se fez foi substituir o uso do juízo humano nas categorizações não revisadas. O problema, claro, é que a necessidade de revisar caso a caso mina o objetivo da aplicação dos modelos, que seria justamente o de evitar a necessidade de analisar cada documento, um por um.

[2] Sobre o COMPAS, ver Brennan e Dieterich (2017).

berdade condicional[2], detectou-se que ela atribui um maior risco de reincidência a pessoas negras (ANGWIN et al., 2016). Foi também detectado viés racial em sistemas de saúde (OBERMEYER et al., 2019), e viés de gênero (desfavorável à mulher) em sistemas de análise financeira (PEACHEY, 2019). Desde que esses e outros casos vieram à tona, teve início uma ampla discussão sobre se e como é possível eliminar, ou ao menos mitigar, esse efeito em modelos computacionais.

Em defesa desses sistemas, há quem argumente que vieses algorítmicos são mais facilmente tratáveis que vieses presentes em juízos humanos: é mais fácil alterar um código do que um coração.[3] Talvez seja verdade, mas isso não significa que eliminar vieses algorítmicos seja fácil. O grau de dificuldade depende, em parte, da transparência com que o modelo aplica os critérios de categorização. Se for possível averiguar em função de quê um dado modelo aplica uma certa categoria a um indivíduo, objeto ou comportamento, será possível submeter essa aplicação à crítica do juízo humano. Contudo, mesmo sendo opacos quanto ao seu funcionamento, esses modelos costumam gozar da confiança das pessoas e isso significa que, muitas vezes, não há sequer motivação suficiente para que se busque averiguar a existência de vieses. Mas é possível, num cenário como esse, justificar essa confiança?

O objetivo desse texto é mostrar que uma boa resposta a essa pergunta depende de uma análise mais profunda e abrangente do que se tem feito até o momento. Para isso, adotou-se a seguinte estratégia: na seção 2, serão introduzidas algumas características gerais dos modelos computacionais, e o modo como vieses podem sedimentar-se. Na seção 3, faremos uma conexão entre problemas clássicos das pesquisas em IA e o desafio de mitigar ou eliminar vieses, concluindo preliminarmente que, embora isso seja possível, é um erro conceber essa tarefa como sendo a de buscar um juízo neutro ou desinteressado. Pelo contrário: um juízo algorítmico não enviesado é precisamente aquele que tem como pano de fundo os valores e interesses da comunidade em que o sistema atua, ou seja, nas regulações que essa comunidade estabelece. Na seção 4, já munidos de ferramental conceitual mais adequado, aprofundaremos a questão da confiança nos modelos computacionais, desenvolvendo a tese de que essa confiança injustificada é parcialmente responsável pela introdução de vieses no modo como as atividades humanas se estruturam e, consequentemente, em nossas tentativas de mitigá-los.

[3] Isso é defendido por Mullainathan (2019).

II

Podemos tomar como ponto de partida a seguinte pergunta: como vieses se estabelecem em modelos computacionais? Uma boa resposta deve considerar a existência de, pelo menos, dois tipos de modelos computacionais: os clássicos e os neurais. Modelos clássicos são aqueles que fazem uso de estatística tradicional. São desenhados diretamente por engenheiros a partir de categorias já conhecidas e com critérios de demarcação claros. Trata-se do tipo de modelo utilizado na maioria dos *softwares* de gestão tradicionais. Neles, as categorizações são feitas por satisfação de regras explícitas que expressam condições necessárias e suficientes. Regras como "será considerado um bom pagador o candidato que não apresentar registros de inadimplência nos últimos 12 meses". Sendo tanto a estruturação dos dados quanto a elaboração das regras uma atribuição dos engenheiros, a presença de vieses será de sua responsabilidade. Vem daí a importância da transparência característica desses modelos: estando as regras explícitas, é possível auditá-las e analisá-las a fim de detectar vieses.[4] A limitação desse tipo de sistema, contudo, aparece já na largada: a necessidade de explicitar todas as regras utilizadas, uma a uma. Em função disso, há uma tendência em sacrificar parte da acurácia e utilizar heurísticas, isto é, regras que podem falhar, mas que capturam um número aceitável dos casos desejados. É perfeitamente possível que bons pagadores tenham enfrentado dificuldades nos últimos 12 meses, e pode-se acrescentar ao sistema algumas regras que tentem reduzir o peso desse critério, caso o candidato receba boa classificação em outros, mas nem sequer se cogita abarcar todas as possíveis variações circunstanciais. Busca-se apenas atingir margens de erro aceitáveis, a depender dos objetivos com o uso do modelo.

O uso de regras hiper-refinadas para aplicar categorias, contudo, é o ponto forte dos modelos neurais. A ideia de utilizar uma arquitetura inspirada na estrutura cerebral é antiga, mas ganhou fôlego renovado na década de 1980 em função do trabalho de McClelland et al. (1987) e Rumelhart et al. (1986). Contudo, o protagonismo contemporâneo desse tipo de modelo nas pesquisas em IA veio apenas muito recentemente, já na década de 2010, numa variante que se convencionou chamar de *deep learning*. Esse protagonismo resultou em uma associação direta entre a construção de modelos neurais e a construção de sistemas inteligentes, como se um sistema pudesse ser considerado inteligente apenas em função de ter sido projetado numa arquitetura neural. Isso é,

[4] É preciso, claro, que o código fonte do sistema esteja disponível para análise, mas o ferramental conceitual necessário para tratar disso, seja na esfera privada, seja na pública, já está bem estabelecido.

no mínimo, impreciso. Apesar disso, não são poucos os produtos comerciais que utilizam variantes do selo "contém IA" (de sistemas jurídicos a escovas de dentes elétricas) graças à popularidade dessa associação. Talvez "contém modelos inspirados em redes neurais" seja menos interessante comercialmente, embora mais preciso. Esse é um ponto importante porque, em larga medida, a confiança que se deposita nesse tipo de sistema vem de sua associação com a IA.

Apesar dessa confiança, redes neurais também apresentam vieses, e como seu uso tem se intensificado nos últimos anos, é importante compreender algumas de suas particularidades, a começar pela estratégia de geração desses modelos. Enquanto modelos clássicos são desenhados diretamente pelo engenheiro, modelos neurais são gerados indiretamente por um algoritmo de treinamento.[5] Esse algoritmo é aplicado sobre uma base de dados e busca, grosso modo, detectar correlações entre os elementos ali presentes. Se a base de dados for composta por, digamos, fotos de elefantes, o resultado poderá ser um modelo capaz de distinguir se existe ou não um elefante numa foto, ainda que essa foto não esteja entre as que foram utilizadas para treiná-lo. O que o algoritmo de treinamento faz é análogo a tentar identificar, ele mesmo, regras ou critérios que podem ser utilizados para detectar a presença de um elefante. Como o processo é automatizado, a limitação que se apresentava nos modelos clássicos é mitigada: modelos neurais são capazes de condensar uma quantidade gigantesca de critérios interdependentes entre si. Ele pode perceber, por exemplo, que a presença de presas de marfim é um elemento relevante, mas não essencial, e o mesmo ocorre com o número de patas, orelhas ou algumas de suas propriedades (cor, tamanho, etc.). O mesmo tipo de procedimento pode ser utilizado para gerar modelos minuciosos da atividade humana, permitindo um juízo bastante refinado, personalizado e sensível a variações circunstanciais.

Sendo assim, modelos neurais podem apresentar vieses tanto em função do algoritmo de aprendizado, quanto em função do conteúdo presente nos dados utilizados para o treinamento. É possível, por exemplo, treinar um modelo que compile os perfis históricos de ocupantes de um determinado cargo e usá-lo para identificar perfis semelhantes em novos candidatos. Nesse cenário, ainda que o gênero não seja um critério explícito, a presença diminuta de mulheres no histórico de ocupantes daquele cargo pode fazer com que o modelo privilegie currículos masculinos. Contudo, seria apressado concluir

[5] É comum que se fale também em "algoritmo de aprendizado", ou "aprendizado de máquina" (*machine learning*). Todos esses termos serão aqui utilizados como tendo o mesmo significado.

que a presença de vieses é sempre resultado direto de um desequilíbrio estatístico. O estudo de Rambachan e Roth (2019), por exemplo, sugere que a relação é menos direta, e que o modelo resultante poderia apresentar, inclusive, o viés inverso, a depender do modo como o treinamento se deu. Isso significa que, para detectar um viés num modelo neural, não é suficiente fazer uma análise dos dados utilizados para o treinamento. Igualmente insuficiente é verificar se o modelo apresenta os resultados esperados contra um conjunto de casos utilizados para testá-lo. Um resultado correto não garante que o "raciocínio" utilizado pelo modelo seja adequado a ponto de podermos confiar nas suas predições em casos novos.

Uma das razões para essa dificuldade é que nem sempre existe uma referência clara contra a qual se possa mensurar a eficiência de um modelo. Caso o objetivo seja detectar elefantes ou identificar pessoas usando máscaras numa multidão, então é possível averiguar o desempenho, afinal, se o sistema confundir um esquilo com um elefante, o erro será evidente. O mesmo não acontece, contudo, quando se faz uso desse tipo de modelo para predizer comportamentos futuros. Como avaliar o grau de eficácia nesses casos? O único tipo de critério disponível depende de contrafactuais ("se lhe houvesse sido concedida liberdade condicional, ele teria cometido um crime, logo, evitamos um crime"), mas esse é o tipo de afirmação que supõe a acurácia do modelo, justamente aquilo que precisaria ter sido demonstrado. Se realmente fizermos uso desse tipo de raciocínio, estaremos ou incorrendo numa falácia (no caso, petição de princípio), ou confiando numa nada objetiva sensação de que, não fosse pelo uso do modelo, os resultados seriam ainda piores.

Por isso, é importante compreender que tipo de estratégia os modelos neurais utilizam em suas categorizações. Em geral, tenta-se lidar com casos novos a partir de uma espécie de generalização. O que o algoritmo de treinamento tenta fazer, de modo simplificado, é encontrar um grau adequado de tolerância à diferença. É preciso que ela seja grande o suficiente para detectar elefantes que tenham características nunca encontradas (na cor da pele ou na textura, por exemplo) e em condições nunca encontradas (perspectiva, distância, iluminação, etc.). Por outro lado, a tolerância não pode ser grande a ponto de fazer o modelo reconhecer elefantes onde não existem, gerando falsos positivos.

Uma das principais formas pelas quais vieses podem se acomodar em modelos neurais, portanto, se dá no modo como a generalização é realizada. Preliminarmente, é importante notar o problema, antecipado por Hubert Dreyfus[6], de que há incontáveis formas a partir das quais os objetos podem ser semelhantes ou distintos. Propriedades físicas como a presença abundante de água ou uma certa temperatura, estão presentes tanto no corpo humano quanto em um copo d'água. A forma pela qual um algoritmo de treinamento toma certas características como relevantes e outras como irrelevantes pode variar enormemente, portanto. Geirhos et al. (2019), por exemplo, descobriram que alguns modelos adquiriram um viés que os fazia atribuir primazia à textura de um objeto na hora de tentar reconhecê-lo, em detrimento de sua forma. Isso ajuda a entender o porquê de uma criança conseguir reconhecer um elefante, ainda que ele se apresente na forma de um rabisco simples, enquanto um modelo treinado somente com fotos tende a falhar. De fato, aquilo que é tomado como relevante pode soar ainda mais estranho do que uma mera propensão a achar que a textura da pele é uma propriedade essencial para determinar a presença de um elefante.[7] Brendel e Bethge (2019) mostraram que, ao tentar identificar um certo tipo de peixe para o qual um modelo fora treinado, um dos critérios era a presença de dedos humanos. Esse tipo de peixe é rotineiramente tomado como um troféu por praticantes de pesca, e a maior parte das fotos utilizadas para o treinamento era de pescadores segurando-os com as mãos.

Como lidar com esse cenário? Uma primeira possibilidade é a de tentar compensar uma tendência com outra: se o modelo tende a priorizar a textura sobre a forma, ou priorizar um gênero sobre o outro, podemos reverter esse desvio introduzindo uma tendência contrária. Em tese, isso pode ser feito tanto por meio de alterações diretas no algoritmo de treinamento, quanto pela seleção cuidadosa das amostras que constituirão a base de dados utilizada. Porém, mesmo nos casos em que isso se mostra viável, não será uma tarefa simples. A compensação de um viés pode resultar na introdução de vários outros, sem que se perceba. Assim, é preciso analisar detalhadamente o modelo[8] e tentar identificar a cadeia de "raciocínio" que ele segue, e essa não é uma tarefa acerca da qual seja possível garantir uma conclusão confiável em todos os casos.

Tudo isso sugere que a imagem vendida por Mullainathan (2019), segundo

[6] Dreyfus expôs um argumento nessa linha na edição de 1992 do seu famoso *What computers still can't do*, cuja primeira versão fora publicada ainda em 1972. Ver a introdução adicionada em Dreyfus (1992).

[7] O exemplo é utilizado por Gary Marcus (2018; 2019) em sua crítica ao *deep learning*.

[8] A análise de agrupamentos (*clusters*) é um exemplo de ferramenta que pode ser utilizada para esse fim.

quem é mais fácil mudar algoritmos do que corações, pode ser enganosamente simplista. Vieses algorítmicos nem sempre podem ser facilmente detectados, tampouco facilmente resolvidos. Dado o que se discutiu até aqui, contudo, não vimos razão para deixar de pensar que se trata de um problema de engenharia. Os resultados questionáveis apresentados por sistemas atuais seriam explicáveis por técnicas falhas ou pouco amadurecidas que podem vir a ser corrigidas. Nessa perspectiva, a eliminação dos vieses é questão de tempo.

Porém, há armadilhas. A principal delas é que tratar de vieses é um problema geral para o qual existem apenas soluções particulares. Para entender a natureza desse problema, pode-se fazer uma analogia com o conceito de relevância: um objeto só se torna relevante no interior de uma tarefa específica. Se o gerente de uma empresa quer repassar uma tarefa a um funcionário e manda-lhe uma mensagem como "preciso que você realize uma tarefa, então vá separando todos os materiais relevantes e, em meia hora chego aí e lhe explico o que preciso", gera-se um problema. Como é possível supor o que será relevante antes de saber qual a tarefa adequada? A relevância não é uma propriedade intrínseca dos materiais em questão, mas sim do papel que eles exercem no interior de algum processo. Esse papel pode variar radicalmente em função dos objetivos, expectativas e especificidades de cada situação. Assim, não parece possível uma teoria geral da relevância que aponte regras gerais como "será relevante todo elemento que. . . ." de um modo independente das situações concretas. Em vez de uma teoria, o resultado seria uma lista sem fim de possíveis tarefas que se possa vir a querer realizar em diferentes situações, indexando cada possível combinação a uma lista dos elementos que lhe seriam relevantes.

Algo análogo se dá no caso dos vieses. Não parece possível criar uma teoria geral do viés, que permita aos engenheiros elaborar uma solução geral para a questão. A depender dos objetivos específicos e das situações concretas com as quais lidamos, o que pode ser adequado num caso pode se mostrar inadequado em outro. Isso sugere que pode haver uma dificuldade mais profunda envolvida na tentativa de gerar modelos isentos de viés. Uma dificuldade que não depende da existência ou não de boas técnicas de engenharia, e que não pode, portanto, ser tomada como um mero problema técnico. Não seria a primeira vez em que uma empreitada científica se depara com desafios análogos. Em particular, problemas de relevância aparentemente insolúveis são velhos conhecidos dos pesquisadores em IA, e o conhecimento acumulado em torno das tentativas (falhas) de solucioná-los se mostra bastante útil.

III

Classicamente, a inteligência foi concebida como algo muito próximo da capacidade para o raciocínio lógico e matemático. Não por acaso, ainda nos anos 1990, quando o sistema *Deep Blue* venceu Kasparov numa partida de xadrez[9], muitos tomaram o evento como uma prova do sucesso da IA e da validade dos seus pressupostos: a inteligência humana pode ser modelada computacionalmente, afinal.[10] Como o *Deep Blue* conseguia antecipar um número de jogadas muito maior do que um ser humano consegue, parecia possível dizer que o computador tinha, de fato, não apenas apresentado um comportamento legitimamente inteligente, mas também que o fez num grau superior ao nosso.

Contemporaneamente, contudo, pode-se afirmar que essa concepção de inteligência abarcava somente a ponta do iceberg. Isso se mostra quando deixamos de prestar atenção aos alegados sucessos da IA, e passamos a focar nas razões dos seus fracassos. Foram eles que tornaram saliente a importância da sensibilidade ao contexto, isto é, a capacidade de se adaptar, de modo rápido e fluido, às mais variadas nuanças circunstanciais. Inicialmente, a IA tentou reproduzir essa capacidade a partir do mesmo ferramental lógico e matemático que utilizava para outros feitos, como o de jogar xadrez. A natureza das dificuldades encontradas, contudo, sugere que algo radicalmente diferente está em jogo: a sensibilidade ao contexto não é um feito da capacidade de raciocínio lógico, mas uma condição necessária para que ela seja possível. Como veremos, na ausência de um contexto bem delineado, qualquer raciocínio tende a se perder em infindáveis inferências completamente irrelevantes. Mas por que a sensibilidade ao contexto é um desafio? Para entender isso, é importante conhecer o modo como esse problema se manifestou na história pregressa da IA.

Toda situação em que uma dada ação ou decisão ocorre pode ser caracterizada, ainda que de modo vago, por um enquadramento (também chamado de *frame*), tal como: estar em casa, estar de férias, realizar uma compra, etc. Quando adentramos o carro de um motorista de aplicativo, por exemplo, o comportamento adequado, tanto de nossa parte, quanto por parte do motorista, depende de compreendermo-nos no interior de um mesmo *frame*. Na IA clássica[11], houve uma tentativa direta de modelar esses *frames* na forma

[9] Em entrevista recente para Fridman (2019), Kasparov confessou ter ficado bastante abalado, visto que aquela não era sua primeira derrota para um computador, mas sim sua primeira derrota em um jogo oficial.

[10] Ver, por exemplo, o debate em Dennett e Dreyfus (2005).

[11] Por "IA clássica" entendemos aqui a empreitada que se desenvolveu com mais ênfase entre os anos 1960 e 1980. Na literatura da área, esse período costuma ser caracterizado pela ênfase no uso de modelos computacionais clássicos, sendo por vezes chamada de IA simbólica ou, seguindo a famosa sugestão de Haugeland (1989), GOFAI (*Good old fashioned AI*).

de conjuntos de regras, mais ou menos como os que constituem e regulam as ações adequadas num jogo de xadrez. Assim como o *Deep Blue* tinha por base um modelo daquilo que é adequado ou não, a depender da disposição de cada peça no tabuleiro, havia a crença de que o mesmo poderia ser feito para guiar comportamentos em restaurantes, escolas e festas. Tais modelos funcionariam como pequenos roteiros que descrevem as estruturas das atividades, ao menos em condições normais.[12] Num restaurante, poderíamos dizer que a entrada é servida primeiro, e que é aceitável rejeitar uma sobremesa, mas que deixar de pedir um prato principal demandaria uma justificativa plausível. Em tese, isso permitiria que um sistema guiado por tais modelos "soubesse" como se comportar naquele ambiente, assim como "sabem" se comportar num jogo de xadrez.

A despeito do otimismo quanto a essa abordagem, logo descobriu-se que jogos de xadrez e restaurantes funcionam de modo bastante distinto. Restaurantes não parecem tratáveis nos termos de um conjunto de regras porque não constituem domínios autônomos, isto é, nosso comportamento em restaurantes não se deve apenas àquilo que sabemos sobre restaurantes. Há uma interação marcante com vários outros domínios, mesmo em condições normais: se estamos num restaurante e encontramos ali um colega de trabalho, uma série de regras ligadas a esse outro domínio pode se fazer valer ali (talvez seja prudente beber menos do que o desejado). O mesmo vale para encontros que tragam más lembranças, uma notícia ruim que chega ao celular ou uma mosca a incomodar.

Apesar disso, seria um erro inferir que restaurantes não constituem um domínio distinto. *Frames* são uma espécie de domínio aberto, isto é, um domínio cujas bordas são borradas, mas não a ponto de se dissolverem: há sim regras que guiam, ainda que de modo vago e geral, nossa compreensão do que caracteriza um comportamento adequado ou inadequado em restaurantes. Contudo, ao contrário do xadrez, essas regras não tratam explicitamente de todas as situações possíveis que podem ocorrer naquele ambiente. No xadrez, podemos dizer algo como: "mover uma torre na transversal é ilegal, e você deve, portanto, voltar atrás nessa jogada". Já num restaurante, não faria sentido dizer: "correr não consta nas regras que guiam os comportamentos possíveis em restaurantes, por isso só me resta ignorar o seu aviso de fogo". Um modo de compreendermos essa diferença é percebendo que toda regra vigente no interior de um *frame* carrega consigo uma cláusula *ceteris pa-*

[12] Entre as abordagens mais famosas estão os *scripts* de Schank (1975) e os *frames* de Minsky (1997). Note que, em Minsky, o termo *frame* é utilizado não para designar enquadramentos que caracterizam estruturas das atividades humanas, mas sim a estrutura de dados utilizada para modelar computacionalmente esses enquadramentos.

ribus. Algo como "proceda de tal modo, a menos que haja uma razão suficientemente forte em contrário". Mas o que conta como uma razão suficientemente forte? Essa questão será mais desenvolvida logo adiante. O importante a notar, por ora, é que a capacidade de seguir regras que contém cláusulas *ceteris paribus* é dependente da capacidade de discernir quando, e em que medida, é adequado segui-las. Acompanhando uma tradição da IA sobre a qual falaremos logo adiante, vamos chamar essa capacidade de *senso comum*.

Podemos agora sintetizar esse ponto afirmando que a distinção entre o domínio do xadrez e o domínio de um restaurante é que um comportamento adequado no interior do segundo depende do senso comum. Como essa dependência emerge? Chamemos os domínios dependentes do senso comum de *domínios abertos* e os independentes de *domínios fechados*. Em domínios fechados, um contexto (uma situação específica no decorrer do jogo, por exemplo) é sempre definido e articulado a partir das regras que constituem aquele domínio. Se um jogador observa o tabuleiro e interpreta a disposição das peças como caracterizando uma ameaça ao seu rei, os únicos fatos relevantes para essa caracterização são aqueles oriundos das regras do jogo. Nesse sentido, o contexto de "ameaça ao rei" é dito *saturado*. Não há nenhum fato novo que possa ser inserido no sistema e que venha a se mostrar relevante na interpretação da atual situação por parte do jogador. Seria possível, por exemplo, informar ao jogador ou acrescentar a um modelo o número de vezes que cada peça foi usada, ou a média de movimentos que cada peça costuma realizar antes de ser tomada, mas nada disso seria relevante para modificar a compreensão de que o jogador (ou o modelo) tem da atual situação. Consequentemente, nenhum destes fatos serão relevantes para determinar o comportamento adequado.

Já em domínios abertos, como no caso dos *frames* que caracterizam as estruturas das atividades humanas, o contexto é *insaturável*. Isso significa que qualquer fato pode se mostrar subitamente relevante na hora de determinar o que constitui um comportamento adequado ou não naquela situação. Vejamos um exemplo para tornar essa ideia mais clara. Suponha-se que, por uma razão qualquer, estejamos interessados em prever se o cachorro de um vizinho latirá no próximo sábado.[13] Seriam as informações acerca do campeonato mundial de xadrez que se realizará sábado, num país distante, relevantes para nossa empreitada? Parece evidente que não. Contudo, bastam dois fatos para que isso se inverta radicalmente: primeiro, que o vizinho em questão é irmão de Magnus Carlsen, atual campeão mundial de xadrez que defenderá

[13] O exemplo é inspirado em Samuels (2010).

seu posto no próximo sábado. Segundo, que ele tem por hábito comemorar as vitórias do irmão com fogos de artifício. Estes dois fatos tornam saliente uma parte de nosso conhecimento que, até aqui, parecia irrelevante: cães tendem a se assustar e latir em função de fogos de artifício. Subitamente, esse conjunto de fatos deixa claro que o campeonato mundial de xadrez aumenta as chances de o cachorro latir no sábado.[14] Esse exemplo ilustra a tese de que, em domínios abertos, não há caracterização prévia de todos os fatores potencialmente relevantes, tampouco do efeito que eles podem vir a ter sobre o comportamento humano.

A distinção entre domínios abertos e fechados é importante para compreender os problemas enfrentados pela IA porque, em modelos computacionais, todo contexto é forçosamente tratado como se fosse saturado. A tentativa de modelar contextos insaturáveis enfrenta desafios que muitos autores entendem ser insuperáveis. Para nossos propósitos, o mais relevante deles é o que ficou conhecido como o problema do senso comum. Já vimos que o senso comum envolve a capacidade de lidar com cláusulas *ceteris paribus*, mas ainda não sabemos exatamente em que isso consiste para as pesquisas em IA. Numa primeira aproximação, o senso comum é tanto um conjunto de conhecimentos acerca da estrutura das atividades humanas (os *frames* que as constituem) quanto um certo modo de fazer uso dele. Modelar computacionalmente essa forma caracteristicamente humana pela qual se faz uso desse conhecimento é um desafio bem conhecido pela IA desde seus primórdios, em 1956.[15] Mccarthy (1968), um dos principais pesquisadores da época, detectou esse desafio e dedicou-se a ele já na largada da empreitada, sem grande sucesso, contudo. Uma das razões do fracasso é que modelar o senso comum envolve tratar de um tipo de conhecimento que nos soa tão evidente, que é difícil até mesmo explicitá-lo.

Um novo exemplo pode ajudar. Suponha-se um cenário em que se está a ensinar uma receita para alguém. A exposição das instruções parece trivial: apontam-se os ingredientes necessários, a quantidade adequada, o tempo que devem permanecer ao fogo, etc. Quão estranho seria, durante a exposição das instruções, ouvir perguntas como essa por parte do aluno: pode a cor do açúcar se alterar se eu tirá-lo do pacote e colocá-lo nesse pote? Se eu encostar a faca na manteiga, pode a faca derreter? Se eu colocar água na jarra de vidro, ela vai retê-la? Notei que você colocou uma faca na gaveta e não a vejo mais, ela ainda existe? Percebi que você deu exatamente 78 voltas ao tentar

[14]Outro fato que se torna saliente é que o vizinho em questão é pouco sensível ao bem-estar canino.
[15]Conforme Boden (2006).

bater os ingredientes, funcionaria também com 77? E com 79?

O conhecimento que evitaria esse tipo de pergunta é o que constitui o senso comum. Note que não se trata de um saber obtido via educação formal, pois espera-se que mesmo uma criança, por mais imaginativa que seja, não tenha dúvidas como essas. Isso sugere que, para que um modelo computacional seja capaz de realizar aquela atividade, ele precisa que figure ali, de modo explícito numa espécie de lista gigantesca, essas e outras incontáveis "obviedades". Sem isso, o sistema não será capaz de raciocinar como um ser humano, dando vazão a comportamentos bizarros.

Elaborar essa lista parece trabalhoso, mas por que acreditar que esse pode ser um problema insolúvel? Um primeiro desafio é o de gerar a informação. Pode-se imaginar que o *Big Data* é de grande valia, mas as informações tipicamente presentes nele raramente servem a esse fim. O que existe ali são dados sobre decisões tomadas e comportamentos adotados, não sobre aquilo que os motivou ou sobre a cadeia de raciocínio que lhes deu origem. É provável que as informações necessárias para evitar as perguntas estranhas do aprendiz de cozinha não existam em base de dados alguma. Parece que dependemos, portanto, de um esforço reflexivo para tentar identificar em larga escala todos os pequenos conhecimentos que caracterizam o senso comum. É uma tarefa colossal, mas há quem tente realizá-la. O exemplo mais lembrado é o CYC de Lenat et al. (1990). Ele iniciou seus trabalhos na década de 1980 e, embora tenha conseguido gerar um produto comercial aplicável a automações, não há nenhum avanço significativo no objetivo mais amplo de modelar o senso comum. Segundo o site da empresa Cycorp (2020), sua base de conhecimento dedicada ao senso comum possui hoje mais de dez mil predicados, 25 milhões de asserções e milhões de coleções de fatos e conceitos. Seria um tanto surpreendente se realmente precisássemos supor a existência de todo esse volume informacional em nosso aparato cognitivo apenas para explicar a capacidade de concluir que uma faca não deixa de existir ao ser guardada.

Acumular informações, no entanto, é a parte fácil. Mesmo um sistema que detenha toda a informação necessária acerca de todos os possíveis contextos, precisa também ser capaz de estruturá-la adequadamente, ou não poderá circunscrever a porção desse saber que é relevante para a situação específica em que se encontra. Como vimos anteriormente, os contextos que caracterizam nossa estruturação do mundo têm caráter insaturável. Isso significa que, aquilo que é adequado ou não a uma dada circunstância pode variar enormemente em função de uma quantidade potencialmente infinita de elementos. No exemplo do restaurante, mesmo fatos sobre elementos ausentes (no caso,

colegas de trabalho) podem ser relevantes para caracterizar o contexto a partir do qual a atividade se dará, e essas possibilidades precisam estar presentes nos modelos, ou o sistema agirá como se não existissem. Mas se esse é o caso, como distinguir, por princípio, ausências relevantes das irrelevantes? Podemos perceber agora a falta que nos faz uma teoria geral da relevância, pois emerge aqui uma dificuldade crucial: um *frame* não pode ser computacionalmente modelado tal como se fosse um *script* ou roteiro vago daquilo que é adequado ou não e daquilo que é relevante ou não em uma dada situação. Isso geraria um comportamento inadequado por parte do sistema, pois a capacidade de seguir um tal *script*, sem se perder na consideração de possibilidades esdrúxulas e irrelevantes, pressupõe o senso comum. Ele é o responsável por resolver as vaguezas e preencher tudo o que não estiver explicitado no roteiro. Para embutir o senso comum em modelos computacionais, resta então uma única alternativa: encontrar uma forma de "catalogar" e organizar todas as possibilidades lógicas, tratando domínios abertos como gigantescos domínios fechados. Essa é a razão pela qual modelos computacionais só conseguem tratar de contextos como se fossem saturados. Não foi por acaso, portanto, que esse caminho tenha sido escolhido pela IA clássica. Era sua única opção.

Sendo esse o caso, como deve esse conhecimento ser organizado? Uma exigência essencial é que essa estrutura seja isomórfica à estrutura de *frames* que caracterizam as atividades humanas. Na história da IA, esse problema organizacional ficou conhecido como *frame problem*.[16] Ele se mostrou um desafio especialmente difícil porque a organização adequada é, ela mesma, dependente de contexto. Repare como a dificuldade do aprendiz de cozinha não se dava pela completa ausência de conhecimento, mas pela sua desorganização. Ele dispunha de inúmeras informações sobre o mundo, mas não conseguia fazer bom uso delas, originando perguntas descabidas no interior daquele *frame*. Espera-se que um ser humano seja capaz de reconhecer que está no interior de um *frame* como "aula de gastronomia", e que guie sua compreensão dos contextos que ali encontrar a partir disso. Diante de uma instrução como "mexa até ficar consistente", é essa capacidade que lhe permitirá entender que o verbo mexer diz respeito ao preparo, e não a um talher (que pode se mexido fora da panela) ou ao seu corpo (que o aprendiz pode mexer, mas sem qualquer efeito sobre o preparo). Essa mesma compreensão lhe permitirá ordenar os conhecimentos, priorizando os mais relevantes: como

[16]Note que, na literatura da IA, o termo "frame" em "frame problem" pode ter significados distintos do que se vem chamando aqui de *frame*, como é o caso em Minsky (1997). Contudo, a natureza do problema é a mesma. Para uma defesa dessa compreensão, ver Barth (2018).

detectar a consistência? Deve-se usar as capacidades motoras, de modo a sentir a crescente resistência do preparo aos movimentos, ou deve-se priorizar o conhecimento matemático, de modo a contar o número de voltas que a colher usada faz? Não é difícil imaginar pequenas variações contextuais que alterariam completamente a prioridade atribuída a cada parcela de conhecimento do senso comum. Assim, o *frame problem* não é apenas o problema de encontrar "o" modo certo de organizar a informação que caracteriza o senso comum, mas sim o problema de como modelar a nossa capacidade de rever continuamente essa organização (e as prioridades atribuídas a cada parcela em função de sua relevância na situação concreta), e isso para todo e qualquer contexto, mesmo aqueles com os quais nunca tivemos contato prévio.

Para entender porque esse é um problema grave, basta notar que essa organização é fruto de nos compreendermos no interior de um dado contexto caracterizado por um dado *frame*: "estou agora numa sala assistindo a uma aula de gastronomia". Mas que tipo de critério pode ser usado para concluir a natureza dos contextos em que nos encontramos? Imagine-se que alguém está descrevendo objetos presentes num cenário e nós tenhamos que adivinhar que cenário é esse. Já sabemos que ali há copos, colheres, mesas, comida, líquidos, alguém servindo e alguém sendo servido. Podemos concluir que se trata de um restaurante? Não. Note-se como todos esses elementos são também compatíveis com quartos de hospitais e mesmo refeitórios de certos presídios. O que fazer diante disso? É suficiente aumentar a quantidade de elementos verificados? Acrescentemos, por exemplo, a informação de que há talheres de plástico. Isso conta pontos a favor da hipótese do refeitório no presídio e do hospital, mas também torna saliente a possibilidade de ser um restaurante com uma área infantil. Quantas características a mais deveríamos checar então, antes de concluirmos com segurança? A depender do contexto, um único elemento saliente pode ser suficiente, enquanto em outros casos, mesmo que chequemos dezenas, a questão permanecerá em aberto. Além disso, a qualquer momento uma nova informação pode nos fazer rever o significado e a relevância de todos os elementos anteriormente considerados, como quando uma investigação policial sofre uma reviravolta em função de novas evidências.

A lição a extrair desse exemplo é que os critérios utilizados para detectar se um certo cenário recai ou não sob uma dada categoria são, eles mesmos, dependentes de contexto. Não há critérios gerais. Para tentar lidar com isso, poderíamos seguir muitos dos pesquisadores da IA clássica e imaginar que *frames* se organizam de forma hierárquica: o *frame* "restaurante" seria uma espécie do gênero "evento social". Um modelo

computacional poderia então proceder segundo essa hierarquia: dado que se trata de um evento social, pode-se descartar a possibilidade de ser um hospital ou um presídio. Contudo, não é bem assim. Essa hierarquia pode representar o que é típico, mas os contextos em que realizamos nossas atividades não estão limitados a essa tipicidade. Emergências hospitalares podem ocorrer em restaurantes, e jantares românticos podem se dar em hospitais. Gera-se, assim, a necessidade de multiplicar indefinidamente o número de *frames* e de variações possíveis destes. O modelo computacional precisaria antever e organizar todas as variações circunstanciais possíveis, bem como todos os caminhos a partir dos quais os contextos se alternam ou se mesclam entre si (um jantar romântico se torna uma emergência hospitalar caso alguém passe mal, e o fato de os dois serem colegas de trabalho pode ter efeitos enormes sobre o que se considera um comportamento adequado nessa situação).

Em síntese, o que temos é um cenário onde parece inescapável que tratemos as situações do mundo humano tal como tratamos jogos de xadrez. O *frame problem* nos limita a modelos que tratam do mundo humano como um gigantesco domínio fechado onde todos os fatos que podem vir a ser relevantes precisam estar previamente explicitados. Isso é desastroso, pois significa que não há como modelar o senso comum sem que disso resultem problemas computacionalmente intratáveis.[17]

Assim caracterizado, o *frame problem* parece associado ao uso de modelos computacionais clássicos. Neles, o desafio toma a forma da tediosa e proibitiva tarefa de elaborar uma gigantesca lista de regras que delineie cada possível contexto humano, bem como suas propriedades e as possíveis relações entre eles. Não por acaso, essa abordagem foi deixada de lado. Mas permanece em aberto a questão sobre se os modelos neurais teriam condições de evitar que o problema venha à tona. Afinal, como vimos, uma de suas principais vantagens era justamente a de se mostrar sensível a um grande número de circunstâncias sem a necessidade de modelar explicitamente um grande conjunto de regras. Vários autores depositaram suas esperanças nisso.[18] Infelizmente, esse tipo de modelo enfrenta sua própria versão dessa dificuldade.

[17] O frame problem foi primeiro descrito por Mccarthy e Hayes (1969) e foi tema de ácidos debates nos anos 1980 e 1990, tendo sido retomado por um pequeno círculo no início dos anos 2010, sem qualquer sucesso ou avanço na sua solução, contudo. Uma antologia desses debates pode ser encontrada em Pylyshyn (1987), Ford e Pylyshyn (1996) e Kiverstein e Wheeler (2012). Contemporaneamente, no campo das ciências cognitivas (psicologia e neurociência), esse tipo de dificuldade é evitada por apelo a elementos não computacionais. Isso originou variantes diversas da chamada *cognição situada*. Ver, por exemplo, Varela, Rosch e Thompson (1992), Clark (1998) e Chemero (2009). Para esses autores, a inteligência humana não pode ser modelada em termos computacionais. No caso da IA, contudo, aceitar isso seria equivalente a desistir.

[18] Ver, por exemplo, Churchland (1989).

Algoritmos de treinamento também enfrentam o desafio de arregimentar e organizar toda a informação necessária para caracterizar o senso comum. O que distingue essa abordagem é o modo pelo qual se busca superar o problema: em vez de tentar modelar a estrutura diretamente, busca-se encontrar a base de dados correta, isto é, aquela que permita ao algoritmo de treinamento inferir os *frames* que deram origem ao comportamento registrado. Em parte, vem daí a ideia comum de que um maior volume de dados pode resolver qualquer problema de aprendizado, afinal, quanto maior a base de treinamento, maiores as chances de que um subconjunto dela seja exatamente o que é preciso para inferir as estruturas subjacentes à atividade em questão. Como vimos, em modelos neurais, inferir a estrutura de *frames* certa (isto é, aquela que de fato originou o comportamento humano registrado) significa encontrar o modo certo de generalizar ou "abstrair" os critérios, regras ou padrões que o guiam.

Um primeiro problema, portanto, é o de que há um número indeterminado de abstrações compatíveis com os dados. Marcus (1998) fornece uma ilustração dessa dificuldade. O autor treinou modelos que deveriam receber um número, processá-lo e apresentar o resultado adequado. A ideia era simples: se a entrada fosse 2, a saída deveria ser 2; se a entrada fosse 6, a saída deveria ser 6, se a entrada fosse 10, a saída deveria ser 10, e assim por diante. A um ser humano, bastam estes exemplos para que se perceba a regra subjacente: $f(x)=x$, ou seja, dado um número qualquer, a resposta certa é esse mesmo número. Contudo, no conjunto de dados utilizado para geração do modelo haviam apenas números pares. Se o modelo gerado conseguir abstrair a regra desejada dos dados, segue-se que, mesmo diante de um número ímpar com o qual ele nunca teve contato prévio, ele não se acanhará e aplicará a mesma regra: se entrada for 3, a saída será 3. Isso não aconteceu. O sistema mostrava um número par qualquer, de forma um tanto errática, deixando claro que ele não conseguiu abstrair a regra adequada, e que estava se guiando por algum outro critério. Isso é possível porque há múltiplas regras que são compatíveis com os dados.

Soma-se a isso a já mencionada dificuldade de compreender que tipo de critério ou padrão está norteando as inferências do modelo: como vimos, não é suficiente analisar os dados nem os resultados. Suponha-se que tentemos corrigir o modelo de Marcus treinando-o a partir de um conjunto de dados que inclui uns poucos números ímpares. Isso significa que ele agirá de acordo com a regra $f(x)=x$? Talvez, mas não necessariamente. O sistema poderá "concluir" que a regra adequada é *"se x é par, a saída será x; se x é ímpar, a saída será x-1"* ou ainda alguma outra regra qualquer,

que pode ser de difícil determinação.[19]

Uma outra versão dessa mesma dificuldade se mostra quando tentamos fazer com que um modelo distinga quais, dentre as regularidades ou correlações mapeadas, constituem relações causais. O papel do senso comum pode ser observado tanto em exemplos de falsos negativos (correlações que não implicam causalidade) quanto de falsos positivos (correlações que, apesar de estranhas, caracterizam relações causais). Vigen (2015) exemplifica que, a considerar os dados entre 1999 e 2009, existe uma correlação entre o número de filmes em que Nicolas Cage aparece e o número de pessoas que morreram afogadas ao cair numa piscina. Dado o senso comum que partilhamos, é desnecessário argumentar que não há uma relação causal entre uma coisa e outra. Mas essa compreensão está fora do alcance de um algoritmo de aprendizado, podendo ele concluir qualquer coisa a respeito, bastando que seja algo compatível com os dados. Por outro lado, há correlações um tanto estranhas que são, de fato, relações causais. É bem conhecida, por exemplo, a relação entre o tamanho da mão de um indivíduo e o tamanho do vocabulário que ele domina em seu idioma natal. Essa relação pode ser caracterizada como causal porque, em geral, crianças tendem a ter mãos e vocabulários menores que adultos.[20]

Historicamente, o problema do senso comum foi evitado pela circunscrição dos sistemas, por parte dos engenheiros, a contextos artificialmente delimitados. Esse é o resultado direto do foco em problemas específicos, tais como o de jogar xadrez, e nada mais. Isso diz muito sobre a natureza do trabalho realizado pelo engenheiro de um sistema: o que ele está fazendo é dando origem a um domínio fechado ou replicando algum já existente.[21] O engenheiro adota um contexto de atividade humana como ponto de partida e "recorta" um conjunto de elementos e regras que serão considerados relevantes. O estado de coisas que não for determinável a partir destes, não fará diferença alguma no comportamento do sistema. No caso do xadrez, o sistema não precisa se preocupar em descobrir se o peso ou a cor das peças é relevante para a próxima jogada, simplesmente porque essas propriedades são inexis-

[19] Como se vê, algoritmos de aprendizado tem pouco apreço pela navalha de Occam.

[20] Esses exemplos ilustram bem uma dificuldade já apontada do *frame problem*: em quais circunstâncias os conhecimentos pertencentes a um *frame* podem afetar conhecimentos pertinentes a outro *frame*? A estruturação do conhecimento que caracteriza o senso comum precisa ser, ao mesmo tempo, rígida o suficiente para sabermos que fatos sobre os filmes de Nicolas Cage não são relevantes para fatos sobre acidentes domésticos com piscinas, mas flexível o suficiente para que percebamos como, às vezes, fatos biológicos de um organismo podem ser relevantes no tratamento de fatos sobre o vocabulário de um indivíduo.

[21] Jogos são bons exemplos de domínios fechados criados por nós. Quando convencionamos um determinado conjunto de regras ou elementos (cartas, peças, tabuleiros etc.) o que estamos fazendo é determinar, por princípio, aquilo que será relevante ou não no interior daquele jogo. Um engenheiro que busque então criar um modelo computacional desse jogo já terá para si um domínio fechado com o qual trabalhar.

tentes nesse domínio. Isso se mostra de forma um pouco mais clara no caso de sistemas que precisem interagir fisicamente com o mundo, pois nestes casos o mundo pode impor ao sistema situações que estão fora daquelas previstas no interior de seu domínio fechado. Note-se como, no caso de robôs em linhas de produção, é preciso alterar e constranger o ambiente em que o sistema será posto para rodar. Se ele foi programado para apanhar um certo tipo de peça e encaixá-la em outra, é preciso que as peças certas estejam nos lugares certos, no momento certo. O robô não precisa tratar de situações em que uma dada peça não seja a que ele espera porque essa possibilidade inexiste nos modelos que ele utiliza para guiar seu comportamento. Uma vez que o mundo não se restringe ao que figura no modelo do robô, contudo, é possível que uma peça não esteja onde deveria estar, mas se isso ocorrer, será uma falha atribuível aos engenheiros que tinham a responsabilidade de adaptar o mundo ao modelo que guia o robô, e não vice-versa. Esse é o caso de qualquer sistema automatizado, em contraste com sistemas inteligentes, dos quais se esperaria a capacidade de se adaptar a situações imprevistas.

Vale notar que esse caráter de sistema automatizado permanece, mesmo que o robô seja extremamente complexo e flexível. Ainda que ele consiga lidar com milhares de peças em dezenas de milhares de diferentes situações (local, peso, forma, etc.), continuamos a tratar de um sistema automatizado, não de um sistema inteligente. A razão para essa recusa em lhe atribuir inteligência precisa estar clara: por mais complexos que sejam, sistemas que operam em domínios fechados só conseguem lidar com contextos saturados e, nesse sentido, dependem que algum agente dotado de senso comum garanta que ele opere num ambiente adequado às suas limitações. Um sistema legitimamente inteligente, como o almejado pela IA, precisa ser independente do senso comum do engenheiro. O sistema precisa ser capaz de descobrir "por si mesmo" o que é apropriado e o que não é, em cada contexto específico, e para isso ele precisa ser capaz de se reconhecer no interior de diferentes *frames*, tal como nós o fazemos. Modelar os *frames* que caracterizam corretamente os contextos de atividade humana é, portanto, condição necessária para que se modele o senso comum, e como vimos, isso passa pela solução do *frame problem*.

O que essa batalha da IA nos ensina é que o senso comum é uma condição necessária para nossa capacidade de raciocínio lógico e matemático, e não um feito dela. A inteligência se demonstra também na própria capacidade de circunscrever domínios abertos em sistemas fechados e estruturados "sob medida" para cada circunstância. Viabiliza-se assim o uso do raciocínio lógico no seu interior sem que nos percamos em infindáveis inferências irre-

levantes.

O que isso significa para a discussão sobre vieses? Nota-se aqui uma razão profunda para que a presença de viés não seja concebida como a presença de perspectivas ou interesses dos agentes envolvidos. Um raciocínio isento de viés não é aquele que ignora tais elementos, atribuindo igual peso a todas as possibilidades lógicas, mas sim aquele que atribui pesos adequados, conforme o contexto, a cada interesse ou expectativa envolvidos. Determinar o peso adequado, como se viu, depende do senso comum, e a tentativa de calculá-los na ausência do senso comum leva ao *frame problem*. Isso não ocorre no caso de domínios fechados suficientemente pequenos, mas dado que tais domínios são, eles mesmos, realizações da inteligência humana, o viés pode se dar já na sua concepção. Aquilo que tomamos como enviesado, portanto, dependerá do modo como nós mesmos circunscrevemos os domínios fechados no interior dos quais queremos automatizar a realização de certas linhas de raciocínio.

Segue-se disso que a mitigação de vieses é possível, mas não é uma tarefa técnica. Não é algo que deva ficar exclusivamente nas mãos de engenheiros. A caracterização de um sistema como enviesado não se dá em função do modo como ele é computacionalmente modelado. Um algoritmo que não implemente um dado domínio fechado de modo correto não é enviesado, mas sim falho. Um *software* de xadrez que movimenta torres na transversal ou cavalos na horizontal não é uma implementação enviesada de xadrez. O programa sequer está jogando xadrez. Essa distinção não é sempre evidente porque, não raro, o engenheiro realiza tanto a estruturação do domínio (o que será ou não relevante, e em que medida) quanto sua modelagem computacional. Nesse sentido, estamos concedendo a eles (ou aos que lhes financiam) um grande poder sobre como nós mesmos devemos estruturar nossas atividades. Ao engenheiro, deve caber somente a responsabilidade sobre a implementação computacional adequada de um dado domínio fechado. A constituição desse domínio, bem como a aceitação ou rejeição de seu uso para um dado fim, são temas de debate público. Deve o algoritmo tratar a todos de modo formalmente idêntico, ou deve ele compensar certas desigualdades materiais? A resposta a esse tipo de pergunta não é de ordem técnica, mas sim política. A mitigação dos vieses, portanto, depende de um pano de fundo regulatório que envolve, necessariamente, o conjunto de valores e interesses públicos vigentes. Tratá-la como uma questão técnica é chegar tarde demais.

IV

Tudo o que foi discutido até aqui torna saliente o quão potencialmente perigoso é o abuso do termo "inteligência"

para caracterizar sistemas automatizados. A recepção que essas tecnologias vêm obtendo nas organizações públicas e privadas raramente leva em conta sua incapacidade de tratar adequadamente o caráter insaturável dos contextos subjacentes às motivações humanas. Isso afeta mesmo tarefas aparentemente simples como a de categorizar textos ou documentos como sendo sobre um determinado tema. Tais tarefas são dependentes da capacidade de compreender linguagens naturais por parte dos modelos, e essa capacidade é dependente do senso comum.[22] Apesar disso, não é raro encontrar instituições que confiem nos resultados apresentados, fazendo deles base para tomada de decisão. Em alguns casos é possível, claro, identificar temas pela presença de certas palavras-chave, mas isso se dá em função de convenções institucionalizadas. Nesses espaços, podemos encontrar regras como "todo documento sobre X deve ter o termo 'X' explicitado em seu cabeçalho". Por si só, a linguagem natural não tem nenhuma demanda análoga a essa. Pode-se escrever um livro inteiro sobre filosofia medieval sem que essa expressão apareça uma vez sequer. Seria igualmente ilusório acreditar que um arranjo complexo de palavras poderá fazer as vezes desses termos chave.[23] Nenhum modelo dentre os atualmente disponíveis seria capaz de compreender, por exemplo, se o presente texto está a extrair lições da história da IA para então aplicá-las a um debate sobre vieses algorítmicos ou se, ao contrário, está a tratar de dificuldades da IA que originam tais vieses.

A confiança rotineiramente depositada na tecnologia, contudo, vai além, e sempre há o risco de um "delírio", fixado no modelo pelo uso de uma estratégia de generalização distinta da utilizada por seres humanos, ser considerado um insight. Como se viu, no caso de modelos neurais, o treinamento pode se dar tanto para fins previamente especificados, como o de gerar um modelo capaz de mapear correlações entre termos de diferentes idiomas (gerando assim um modelo que busca simular a atividade humana da tradução) quanto com objetivos menos definidos, como o de descobrir correlações ocultas entre os dados, quaisquer que sejam. Essa tentativa de descobrir relações não previamente supostas é denominada *data mining* (mineração de dados). Ela permite a geração de modelos que "descubram" novas categorias.

Um exemplo clássico é o de fazer uso

[22] Um exemplo é o GPT-3 da OpenAI, que se tornou famoso por sua capacidade de elaborar textos estruturalmente complexos a partir da "leitura" de documentos ou livros. Contudo, tais modelos se baseiam apenas em análises estatísticas que correlacionam palavras umas às outras. Eles não fazem a menor ideia do que estão falando e são alvos perfeitos para o argumento da sala chinesa exposto por Searle (1980). Para a descrição do GPT-3, ver Brown et al. (2020). Para uma crítica de suas capacidades e bons exemplos dos erros crassos que essa técnica costuma gerar, ver Marcus (2020) e Marcus e Davis (2020).

[23] Exemplos comuns são estratégias como *word2vec* ou *doc2vec*, que representam estruturas sintáticas em vetores numéricos, típicos dos modelos neurais. Isso permite uma comparação refinada da forma de um texto, mas ignora seu significado.

desses modelos para aprender padrões de compra. Ele pode trazer à tona uma correlação como: clientes que gostam de amendoim tendem a preferir cerveja de trigo. Isso origina uma nova categoria, a dos clientes-que-gostam-de-amendoim-e-cerveja-de-trigo, que pode ser utilizada para formular estratégias comerciais a partir de uma hipersegmentação mercadológica numa escala sem precedentes.[24] Diante disso, a pergunta que nos interessa agora é: por que confiamos tão prontamente nas correlações indicadas por esses modelos, tomando-as como descobertas?

Essa confiança é exemplificada pelo artigo de Joshua Davis na *Wired* (2017), onde ele explicita sua expectativa de em breve poder substituir o presidente do seu país (EUA) por uma IA. Para Davis, um tal sistema seria capaz de emitir juízos menos enviesados e de descobrir fatos que nos são inacessíveis, sendo assim capaz de conhecer aquilo que queremos e precisamos melhor do que nós mesmos. Nessa discussão, já vimos algumas razões para sermos reticentes quanto a esse alegado privilégio epistêmico. Na ausência do senso comum, os modelos são incapazes de distinguir entre correlações espúrias e correlações que caracterizam legítimas relações causais, por exemplo. Na maioria dos casos, contudo, o custo a pagar pela confiança em eventuais erros é relativamente baixo: caso se posicione o amendoim próximo à cerveja de trigo e não se observe um aumento no consumo, nenhum grande dano foi feito. Contudo, se concedêssemos a um modelo o poder que Davis sugere, o potencial para dano seria colossal. A má notícia é que, embora não cogitemos (ainda?) a sério atribuir poder político a um sistema de IA de modo tão direto, estamos sim permitindo que eles regulem nossas vidas em larga escala, e isso está relacionado aos efeitos dos vieses algorítmicos.

Até o momento, tratamos dos vieses como um desvio sistemático em relação a um pano de fundo que caracteriza o que se considera adequado num domínio. Contudo, há um segundo sentido, talvez mais profundo, em que eles podem se mostrar: algoritmos podem introduzir vieses no modo como nós mesmos compreendemos a estrutura de nossas atividades, isto é, nossos *frames*. Nesse caso, os vieses algorítmicos não seriam apenas desvios sistemáticos que ocorrem em modelos computacionais, mas sim um fenômeno em que nossa própria compreensão do que é adequado ou não estaria sob sua influência, afetando assim nosso senso comum. Como isso é possível?

Esse é um dos efeitos daquilo que Rouvroy (2019; 2015) denominou *governamentalidade algorítmica* (GA).

[24]Importante notar que novas categorias podem surgir mesmo na geração de modelos não destinados ao *data mining*. Os modelos podem gerar categorias que funcionam como intermediárias para os fins estabelecidos pelos engenheiros (a classe das palavras-que-contém-acentos-nos-dois-idiomas para os quais se efetua uma tradução, por exemplo).

Trata-se de uma forma de governo, entendido aqui como um modo de lidar com as incertezas associadas à conduta dos governados, produzindo assim certas regularidades em torno do que se considera adequado ou desejável. Tal regime está presente em várias plataformas informacionais, tanto as puramente virtuais (sistemas de busca) quanto as híbridas (aplicativos de transporte) e tende a transbordar rapidamente para outras dimensões da vida, tais como a política, o direito e a saúde. Rouvroy descreve a GA como fruto da articulação de três práticas específicas: primeiro, a extração capilarizada de informações sobre atividades humanas. Segundo, o uso das informações para gerar modelos computacionais dessas atividades. Terceiro, o uso desses modelos na geração de perfis individuais usados para intermediar as relações destes indivíduos (nós), com instituições diversas, virtuais ou não.

A primeira e segunda práticas já foram aqui desenvolvidas: trata-se tão somente da propensão a registrar cada vez mais aspectos de cada vez mais atividades que realizamos, e de gerar modelos computacionais a partir destes dados. A terceira prática, contudo, é mais específica: não se trata apenas de gerar perfis comportamentais individuais e tentar realizar predições a partir deles, mas de usá-los para afetar ou modular o comportamento individual via mediação das relações entre indivíduos e instituições. É o que se vê, por exemplo, na relação entre indivíduos e empresas como a Google, Facebook e Amazon. Nelas, o perfil individual de cada usuário é levado em conta na hora de apresentar resultados das buscas ou para decidir o que será apresentado em destaque, numa tentativa de predizer a relevância do conteúdo em função dos interesses do usuário. Essa prática fez com que, paulatinamente, a ideia da internet como um oceano a ser desbravado de igual modo por todos, fosse substituída pela imagem de uma lagoa artificial construída sob medida para cada indivíduo. A mesma estratégia pode ser adotada, claro, na venda de bens e serviços em espaços não virtuais. É o que fazem hoje algumas redes de farmácias que apresentam ofertas específicas (não necessariamente vantajosas) em função do perfil de cada comprador.

De que modo o uso dessas tecnologias pode caracterizar um tipo de governamentalidade? A força da GA, nas palavras de Rouvroy (2019, p. 33), está em "*... separar os sujeitos de sua capacidade de fazer ou não fazer certas coisas*". Não há ali oferta de motivações para sugerir um determinado curso de ação. Os filtros são aplicados de modo a mitigar, ou mesmo impedir, as chances de que um dado comportamento venha a ser. Nas palavras de Rouvroy:

> A governamentalidade algorítmica, portanto, apresenta uma nova estratégia de gerencia-

> mento da incerteza que consiste em minimizar a incerteza associada à agência humana: a capacidade dos seres humanos de fazer ou não fazer tudo aquilo que são fisicamente capazes de fazer. Efetuado através da reconfiguração de arquiteturas informacionais e físicas e/ou ambientes dentro dos quais certas coisas se tornam impossíveis ou impensáveis, e lançando alertas ou estímulos produzindo respostas reflexas ao invés de interpretação e reflexão; isso afeta indivíduos em sua agência que é, em sua inatualidade, dimensão virtual da potencialidade e da espontaneidade (...) (ROUVROY, 2019, p. 34)

Exemplos de uso dessa estratégia abundam, pois ela é aplicada em praticamente toda plataforma que objetiva fornecer conteúdo personalizado: plataformas de notícias, áudio ou vídeo que organizam as opções em função do que consideram que será mais interessante são o exemplo mais comum. Também são exemplos os sistemas que almejam fornecer "sugestões" para processos de tomada de decisão, apresentando apenas um subconjunto dos caminhos disponíveis por considerar os demais indesejáveis ou irrelevantes. Todos são casos de reconfiguração no modo como as diferentes possibilidades de ação são apresentadas a cada indivíduo a partir do seu perfil.

Tal estratégia é utilizada não apenas para personalizar o ambiente, contudo, mas também para fazer valer regras de conduta. Nesses espaços, como em quaisquer outros, demanda-se algum tipo de regulação ou controle acerca do que se considera um comportamento desejável. Na medida em que crescem, o volume de casos em que esse controle precisa ser exercido se amplia. Ao fazer uso das estratégias de personalização com esse fim, a GA não atua por meio de sanções ou de caracterizações deontológicas daquilo que deve ser observado. Não se estipula o que deve ou não ser feito, mas sim o que aparece ou não ao sujeito como uma ação possível ou adequada. O contraste com outros regimes é claro: neles, o sujeito compreende que tem diante de si um sem número de possibilidades de ação, mas também compreende que, a depender do caminho adotado, ele estará sujeito a consequências de ordem social ou penal. Uma vez que o conjunto de comportamentos possíveis extrapola o conjunto de comportamentos desejáveis, existem sanções ou orientações. Em um regime de GA, contudo, tais possibilidades são afastadas, ou mesmo negadas ao sujeito, e o espaço dos comportamentos possíveis tende a ser coextensivo com o espaço dos comportamentos aceitáveis.

É fundamental notar como toda essa articulação ocorre de modo invisível ao usuário. O modo como o perfil media

a relação entre o indivíduo e os espaços que ele ocupa, bem como as instituições com as quais se relaciona, ocorre de modo inescrutável. Não se trata apenas de não sabermos que tipo de inferência foi feita a nosso respeito a partir dos nossos rastros, mas também de não sabermos quando ou em que medida essas inferências estão sendo usadas para afetar aquilo que nos aparece como possibilidade de ação. Temos assim um exercício de poder que atua direta e acriticamente nos porões de nossa subjetividade: nosso campo de ação é delimitado e organizado por um processo invisível de ordenação de preferências. Cursos de ação são tornados salientes, ou ocultados, sem que nos seja dada a oportunidade de reflexão sobre as razões que guiaram essa ordenação.

Por que permitimos, afinal, esse tipo de interferência em nosso horizonte de ações? Não se busca aqui dar uma resposta ampla a essa questão, mas apenas enfatizar que uma boa resposta precisa levar em conta o injustificado privilégio epistêmico atribuído aos modelos utilizados na GA, rotineiramente qualificados como "inteligentes". Embora a GA não busque replicar a inteligência humana, mas apenas detectar novas categorias e utilizá-las para fazer segmentações de mercado hiper-refinadas, a ausência de senso comum nos modelos deveria fazer com que tratássemos suas categorizações como hipóteses a serem verificadas, não como descobertas.[25] Sem essa verificação, não temos como saber se o sistema está se deixando guiar por correlações espúrias, por exemplo. Esse é o erro de Davis. Estamos aceitando como verdadeiras quaisquer correlações e categorias que estes sistemas apresentem, sem verificação ou confirmação. Os modelos são incapazes de compreender nossas motivações, e por isso não se pode dizer que nos conhecem melhor do que nós mesmos, mas pode-se sim dizer que estamos deixando que eles nos guiem como se conhecessem.

Diante desse argumento, um defensor da GA poderia insistir: se a ausência de senso comum fosse mesmo um entrave, essa dificuldade se manifestaria na forma de resultados erráticos ou de falhas grotescas nas predições que os modelos fazem a nosso respeito. E não parece ser esse o caso. Muitos modelos usados em ambientes de GA aparentam ter a capacidade de realizar predições acuradas sobre nosso comportamento. Como isso é possível? A explicação aqui oferecida para esse aparente sucesso é a de que estamos diante de uma forma de reflexividade presente no que Hacking (1996) compreendeu como *tipos humanos*.

[25] A rigor, o senso comum é necessário, mas não suficiente. O senso comum permitiria compreender o que está em jogo, e isso poderia viabilizar alguns feitos impressionantes como solicitar a uma IA que leia um conjunto de alguns milhares de artigos e exponha as ideias que guardam relação com nossas pesquisas. Permaneceria em aberto, contudo, em que medida essa compreensão seria confiável a ponto de justificar um privilégio epistêmico por parte do sistema.

Esses tipos (categorias, no jargão que utilizamos até aqui), quando aplicados a agentes humanos, tem caráter reflexivo: o fato de sermos classificados de uma determinada forma influencia o modo como nos comportamos. Para Hacking, esse efeito é o que distingue tipos naturais (polietileno, polipropileno) de tipos humanos (adolescente, refugiado). Quando uma pessoa é tipificada como um adolescente ou um refugiado, altera-se o modo como aquele indivíduo compreende a si mesmo. Em geral, isso se deve a conotações morais associadas ao tipo em questão: o indivíduo pode, por exemplo, não aceitar compreender a si mesmo como um exemplar daquele tipo de pessoa. Essa manifestação pressupõe que o indivíduo esteja, em algum nível, ciente de ser classificado de um tal modo, e que tenha condições de reagir a essa classificação (aceitá-la, criticá-la, rejeitá-la, etc.). Os sistemas utilizados na GA podem dar vazão a esse efeito justamente por serem vistos como fontes confiáveis de novas categorias.

Tome-se um exemplo citado em reportagem da revista *The Economist* (2013): uma determinada empresa mapeou uma correlação estatística entre uma boa performance profissional, bem como uma maior estabilidade no emprego, e o uso de navegadores que não vêm instalados por padrão nos sistemas operacionais[26] para preencher formulários de candidatura para vagas online. A empresa tratou dessa informação como uma descoberta, e seus processos passaram a levar esse dado em conta na hora de decidir quem deve ou não ser contratado. Ou seja, as "descobertas" que os processos de geração de modelos para a GA fazem estão sendo utilizadas para alterar os *frames* em que a atividade humana se dá. Aquilo que se considera relevante para o correto exercício de uma atividade foi revisto. Isso é suficiente para gerar uma pressão normativa. O conhecimento dessa relação (entre performance e uso de certo navegador), pode ser cobrado como algo que bons profissionais de RH devem levar em conta. Candidatos, por sua vez, sabendo que serão "julgados" por um sistema que leva isso em consideração, podem se sentir inclinados a fazer uso de um determinado navegador, a fim de aumentar suas chances.

Há ainda outras formas de manifestação da reflexividade que podem ser associadas ao que Hacking denomina *tipos inacessíveis*: nesses casos, o efeito sobre o agente classificado se dá de modo indireto, por meio de alterações no seu ambiente. O indivíduo não sabe que seu horizonte de ações está sendo afetado em função de uma dada categorização. A reflexividade opera então pelo modo como o indivíduo é tratado pelas pessoas ou pelas instituições em seu entorno. Ora, como se viu, esse é preci-

[26] É o caso do Firefox, em contraste com o Edge, em computadores com Windows.

samente o modo pelo qual a GA opera: alterando o horizonte de possibilidades de ação do indivíduo, ou seja, alterando o modo como o mundo lhe aparece e, consequentemente, o modo como ele se comportará.

Um exemplo de mecanismo psicológico por meio dos quais essa interferência no ambiente pode se dar, são aqueles responsáveis pelos efeitos de *priming*. Como demonstraram Hasher, Goldstein e Toppino (1977), a simples repetição é suficiente: o fato de algum conteúdo aparecer reiteradamente como disponível aumenta as chances de que ele venha a figurar em nossos interesses. Além disso, são conhecidos os efeitos que se dão por associação (pensar em gatos tende a tornar cachorros mais salientes, aumentando a chance de que venham a figurar no raciocínio) ou por categorização semântica (pensar em lobos pode aumentar as chances de que cachorros venham a figurar no pensamento).

Diante da reflexividade das categorias humanas, emerge uma questão crucial: como distinguir uma predição bem-sucedida de uma incitação bem-sucedida? Aquilo que aparece como uma crescente capacidade de predizer uma preferência pode ser, na verdade, uma crescente capacidade de incitá-la, propositadamente ou não. Assim, diante de um modelo computacional cujo uso vem apresentando resultados satisfatórios para a GA, não sabemos por certo se estamos lidando com uma aplicação bem-sucedida de estatística ou com um efeito da reflexividade. No primeiro caso, teremos um modelo que reproduz adequadamente um dado contexto de atividade humana. No segundo, teremos um modelo que modificou com sucesso os contextos nos quais essa atividade é realizada. Talvez a Netflix (ou a Google, ou o Facebook, etc.) não esteja nos conhecendo cada vez melhor, mas nós estejamos nos restringindo, sem perceber e com crescente ênfase, àquilo que o perfil que intermedeia nossa relação com ela nos prescreve. Não surpreende, portanto, a sensação de que as sugestões de conteúdo oferecidas pelas plataformas nos pareçam cada vez mais certeiras.

V

O cenário exposto sugere que o desafio de mitigar vieses está intimamente conectado a uma questão maior sobre onde, quando e como queremos fazer uso de modelos computacionais. Por sua vez, essa questão é profundamente influenciada pela confiança depositada nas promessas das tecnologias envolvidas. Como nos lembra Pariser (2011), essa esperança tem mitigado o questionamento e a demanda pela responsabilização das instituições que fazem uso delas. Na medida em que confiamos a solução dos problemas a esforços de engenharia, tendemos a deixar de lado o fato de que o uso da tecno-

logia é norteado por interesses privados. Consequentemente, por vezes, não há sequer motivação para tratar de problemas mais fundamentais. Tanto predições quanto incitações, por exemplo, são compatíveis com o modelo de negócios que sustenta o uso dos sistemas computacionais. Se deixadas em paz, a tendência é que nenhuma das plataformas tome essa indistinção como um problema. Essa é uma preocupação que não emerge espontaneamente de motivações mercadológicas.

Contudo, como vimos, há um problema ainda mais difícil envolvido. Ainda que se dispusessem a lidar adequadamente com a questão da reflexividade por pressão externa, as plataformas teriam diante de si o problema do senso comum e o *frame problem*, e não está claro nem sequer se é possível solucioná-los computacionalmente. Uma vez que os modelos de IA não apresentam senso comum, a probabilidade de que, por pura sorte, estejam interagindo adequadamente com nossas motivações, é desprezível. Muito mais provável é que estejam afetando as estruturas de *frames* subjacentes às nossas atividades. Por isso, quando atribuímos a eles um privilégio epistêmico injustificado e um poder descabido, permitimos que afetem nossa compreensão do mundo, e que os seus vieses e valores privados embutidos tenham impacto sobre o nosso senso comum. A reflexividade das categorias humanas acaba potencializando esse efeito, pois ela faz com que essa influência transborde para fora das plataformas, em espaços ainda não colonizados por sistemas informacionais: mesmo dicas de conteúdo dadas por amigos trarão consigo os efeitos das plataformas sobre eles.

Em síntese, por um lado, não é razoável exigir das plataformas a solução de problemas fundacionais da IA. Elas não são centros de pesquisa dedicados à questão, mas sim empresas que fazem uso da tecnologia para oferecer serviços. Por outro lado, é igualmente desarrazoado ignorar que o uso corrente dessas tecnologias tem efeitos sociais que não podem ser resolvidos ou mitigados por mais tecnologia. Se queremos lidar adequadamente com os vieses e outros efeitos nocivos, portanto, precisamos revogar o estatuto epistêmico privilegiado tipicamente atribuído aos modelos computacionais. Isso não significa, claro, uma rejeição do uso dos modelos para realizar descobertas. O que precisamos é tomar as inferências por eles realizadas, bem como as categorias por eles detectadas, não como *insights* em seu valor de face, e sim como hipóteses a serem validadas no interior de uma empreitada científica. Além disso, é fundamental compreender os modelos como aquilo que são: automações de atividades em domínios fechados. Desse modo, não deixamos de enxergar que a caracterização desses domínios, isto é, a determinação daquilo que deve ser levado em conta no "raci-

ocínio" realizado no interior dos modelos, precisa ficar sob nossa responsabilidade, e não submetida ao arbítrio dos engenheiros ou das plataformas que os financiam. Na ausência desse posicionamento, a concessão de poder e espaço para sistemas de IA em geral, e para a GA em particular, está sendo facilitada sem que nos apercebamos das consequências, e isso dificulta a discussão sobre possíveis formas de emancipação (se assim desejarmos) ou sobre o que caracteriza uma regulação adequada dos espaços informacionais.

Referências

ANGWIN, J. et al. Machine Bias. *ProPublica*, maio 2016. Disponível em <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Acesso em: 2 jun. 2020.

BARTH, C. *O Frame Problem: a sensibilidade ao contexto como um desafio para teorias representacionais da mente*. Dissertação (Mestrado em Filosofia) – Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, Belo Horizonte, 2018.

BODEN, M. *Mind As Machine: A History of Cognitive Science Two-Volume Set*. Clarendon Press, 2006.

BRENDEL, W.; BETHGE, M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *arXiv preprint arXiv:1904.00760*, 2019. Disponível em: <https://arxiv.org/abs/1904.00760>. Acesso em: 13 jun. 2020.

BRENNAN, T.; DIETERICH, W. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In: *Handbook of Recidivism Risk/Needs Assessment Tools*. John Wiley Sons, pp. 49–75, 2017.

BROWN, T. B. et al. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>. Acesso em: 13 jun. 2020.

CHEMERO, A. *Radical embodied cognitive science*. Cambridge, MA: MIT Press, 2009.

CHURCHLAND, P. M. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge: The MIT Press, 1989.

CLARK, A. *Being There: Putting Brain, Body, and World Together Again*. Cambridge: The MIT Press, 1998.

CYCORP. Cyc's knowledge base. *Cycorp*, 2020. Disponível em: <https://www.cyc.com/archives/service/cyc-knowledge-base>. Acesso em: 7 ago. 2020.

DAVIS, J. Hear me out: let's ellect an AI as president. *Wired*, maio 2017. Disponível em: <https://www.wired.com/2017/05/hear-lets-elect-ai-president/>. Acesso em: 29 jun. 2017.

DENNETT, D.; DREYFUS, H. Did Deep Blue's win over Kasparov prove that Artificial Intelligence has succeeded? In: FRANCHI, Stefano; GUZELDERE, Guven (Eds.). *Mechanical Bodies, Computational Minds*. The MIT Press, pp. 265–279, 2005.

DREYFUS, H. *What Computers Still Can't Do*. The MIT Press, 1992.

ROBOT Recruiters. *The Economist*, abr. 2013. Disponível em: <https://www.economist.com/business/2013/04/06/robot-recruiters>. Acesso em: 16 nov. 2019

FORD, K. M.; PYLYSHYN, Z. W. (EDS.). *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Norwood, NJ, USA: Ablex Publishing Corp., 1996.

FRIDMAN, L. Garry Kasparov: Chess, Deep Blue, AI, and Putin. *Artificial Intelligence Podcast*, out. 2019. Disponível em: <https://lexfridman.com/garry-kasparov/>. Acesso em: 11 set. 2020.

GEIRHOS, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2019. Disponível em: <https://arxiv.org/abs/1811.12231>. Acesso em: 30 nov. 2019.

HACKING, I. The looping effects of human kinds. In: *Causal Cognition*. Oxford University Press, pp. 351–383, 1996.

HASHER, L.; GOLDSTEIN, D.; TOPPINO, T. Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, v. 16, n. 1, pp. 107–112, fev. 1977.

HAUGELAND, J. *Artificial Intelligence: The Very Idea*. Cambridge: The MIT Press, 1989.

KIVERSTEIN, J.; WHEELER, M. *Heidegger and cognitive science*. New York: Palgrave Macmillan, 2012.

LENAT, D. B. et al. Cyc: toward programs with common sense. *Communications of the ACM*, v. 33, n. 8, p. 30–49, ago. 1990.

MARCUS, G. Deep Learning: A Critical Appraisal. *arXiv preprint arXiv:1801.00631*, jan. 2018. Disponível em: <https://arxiv.org/abs/1801.00631>. Acesso em: 21 jan. 2019.

MARCUS, G. GPT-2 and the Nature of Intelligence. *The Gradient*, jan. 2020. Disponível em: <https://thegradient.pub/gpt2-and-the-nature-of-intelligence/>. Acesso em: 26 jun. 2020.
MARCUS, G.; DAVIS, E. *Rebooting AI*. New York: Pantheon, 2019.
MARCUS, G.; DAVIS, E. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*, ago. 2020. Disponível em: <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>. Acesso em: 25 set. 2020.
MARCUS, G. F. Rethinking Eliminative Connectionism. *Cognitive Psychology*, v. 37, n. 3, 1998.
MCCARTHY, J. Programs with common-sense. In: MINSKY, Marvin (Ed.). *Semantic information processing*. Cambridge: The MIT Press, pp. 403–418, 1968.
MCCARTHY, J.; HAYES, P. J. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, v. 4, pp. 463–502, 1969.
MCCLELLAND, J. L. et al. *Parallel Distributed Processing, Vol. 2: Psychological and Biological Models*. Cambridge: The MIT press, 1987.
MINSKY, M. A framework for representing knowledge. In: HAUGELAND, J. (Ed.). *Mind design II: phylosophy, psychology, artificial intelligence*. Cambridge: The MIT Press, pp. 111–142, 1997.
MULLAINATHAN, S. Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times*, dez. 2019. Disponível em: <https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html>. Acesso em: 29 jul. 2020.
OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, v. 366, n. 6464, pp. 447–453, 2019.
PARISER, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. New York: Penguin Publishing Group, 2011.
PEACHEY, K. Sexist and biased? How credit firms make decisions. *BBC News*, nov. 2019. Disponível em: <https://www.bbc.com/news/business-50432634>. Acesso em: 13 set. 2020.
PYLYSHYN, Z. W. (ED.). *The Robots Dilemma: The Frame Problem in Artificial Intelligence*. New Jersey: Ablex, 1987.
RAMBACHAN, A.; ROTH, J. Bias In, Bias Out? Evaluating the Folk Wisdom. *ArXiv preprint arXiv:1909.08518*, fev. 2020. Disponível em: <https://arxiv.org/abs/1909.08518>. Acesso em : 27 ago. 2020.
ROUVROY, A. O(s) fim(ns) da crítica: behaviorismo de dados versus devido processo. In: ALVES, Marco Antônio Sousa; NOBRE, Márcio Rimet (Eds.). *A sociedade da informação em questão*. Belo Horizonte: D'Placido, 2019.
ROUVROY, A.; BERNS, T. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação? *Revista Eco Pós*, v. 18, n. 2, pp. 35–56, 2015.
RUMELHART, D. E. et al. *Parallel Distributed Processing, Vol. 1: Explorations in the Microstructure of Cognition: Foundations*. Cambridge, MA: MIT press, 1986.
SAMUELS, R. Classical computationalism and the many problems of cognitive relevance. *Studies in History and Philosophy of Science*, v. 41, n. 3, pp. 280–293, 2010.
SCHANK, R. C. Using knowledge to understand. *Proceedings of the 1975 workshop on Theoretical issues in natural language processing - TINLAP 75*. Association for Computational Linguistics, 1975
SEARLE, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, v. 3, n. 03, set. 1980.
VARELA, F. J.; ROSCH, E.; THOMPSON, E. T. *The Embodied Mind*. Cambridge: The MIT Press, 1992.
VIGEN, T. *Spurious correlations*. New York: Hachette Books, 2015.

Recebido: 28/09/2020
Aprovado: 18/01/2021
Publicado: 31/01/2021