VIÉS: A nova fronteira de Inteligência Artificial

A intensificação do uso de inteligência artificial (IA) aumentou a preocupação com os vieses (ou bias em inglês), termo usado em estatística para expressar o erro sistemático ou tendenciosidade. As consequências do comportamento tendencioso dos algoritmos são vastas e aplicadas em diversos campos. Isto porque, atualmente, as máquinas tomam decisões e avaliam cenários de forma assertiva e mais rapidamente.

Contudo, os vieses podem ser prejudiciais e afetar vidas de forma definitiva. Uma seleção de pessoal por meio de ferramenta para recrutamento baseada em IA pode, por exemplo, privilegiar alguns em detrimento de outros; IA aplicada para definir sentenças em processos jurídicos também poderia apresentar o mesmo problema, estando mais propensa a sentenciar alguns grupos em detrimento de outros por critérios socioeconômicos, raciais ou de gêneros.

Em um contexto mais de negócios, os vieses podem gerar bolhas e dificultar a inclusão de produtos ou indicações. “A solução de IA indica produtos em um e-commerce ou a lista de recomendação de vídeos em sites baseado no perfil de consumo, o que gera uma bolha, que é consequência de um viés pessoal, porque a máquina reflete o seu próprio viés”, diz Leandro Nunes de Castro Silva, coordenador de desenvolvimento e inovação da Universidade Presbiteriana Mackenzie.

Contudo, os comportamentos tendenciosos são consequências. A principal causa do viés está no dado usado para treinar o sistema, uma vez que o software de IA aprende a partir de alguma experiência passada ou interação com o sistema. “Para treinar um sistema de inteligência artificial para pousar um avião, você vai pegar dados históricos de pousos em diferentes contextos, vai apresentá-los para máquina e mostrar o jeito para pousar.

Com o passar do tempo, a máquina vai absorver aquele conhecimento de como executar a tarefa e vai passar a fazê-la de forma autônoma. É neste ponto que está uma das principais causas de bias. Se você apresentar para máquina dado enviesado, ela vai ficar enviesada. A escolha dos dados e do processo de treinamento são elementos centrais para evitar o viés das máquinas”, destacou Silva.

Assim, um dos desafios está em criar máquinas e sistemas não enviesados. Mas como fazer isto quando as pessoas, que acabam selecionando os dados ou programando as ferramentas, são, por natureza, enviesadas? Uma das maneiras de se mitigar o enviesamento é trabalhar com amostras diversas e envolver no desenvolvimento profissionais de vários perfis. “O enviesamento humano, às vezes, é imperceptível. Então, quando se pegam diferentes fontes, consegue-se mitiga um pouco o viés. Um dos aspectos que toca bastante o viés de máquina é o ético”, pontua o especialista do Mackenzie.

Ainda que a discussão da ética na inteligência artificial seja antiga, o assunto ganhou, recentemente, muita visibilidade, basicamente porque IA conquistou mais espaço como ferramenta de automação. Os impactos éticos do viés têm sido um tema cada vez mais debatido. Qualquer atividade automatizada com inteligência artificial, se estiver enviesada, pode trazer sérios problemas de credibilidade, confiança e reputação.

“Ética, transparência e confiança são a base para a construção de uma IA justa que irá servir à sociedade”, diz Fabrício Lira, executivo de dados e IA da IBM Brasil. Um estudo conduzido pela IBM

denominado “From Roadblock to Scale: The Global Sprint Towards AI” entrevistou 4514 executivos dos EUA, Europa e China, em outubro de 2019, para entender os inibidores em escalar o uso de IA em suas empresas e 78% deles responderam que confiança que seus sistemas de IA produzem resultados justos, seguros e confiáveis é um fator crítico para expandir o uso, enquanto 83% responderam que é universalmente importante saber explicar como a IA chegou a determinada decisão/julgamento.

Por que existe viés

Fabrício explica que pode haver diferentes causas para os vieses existirem quando se fala de inteligência artificial, sendo, usualmente, duas linhas predominantes. “A primeira delas argumenta que o problema reside na demografia dos times envolvidos na criação dos algoritmos, não contemplando um equilíbrio de gênero e racial. A segunda está ligada à existência de viés nos conjuntos de dados utilizados para treinar os algoritmos, o que pode ocorrer de forma acidental ou por questões históricas, carregando vieses raciais, de gênero ou ideológicos”, detalha.

Os algoritmos de IA são cada vez mais usados para ajudar os profissionais a tomar decisões em áreas como medicina, recursos humanos, justiça, varejo e finanças. Nesse contexto, eles podem reproduzir os vieses contidos nos dados em que são treinados.

Contudo, diz Lira, os conjuntos de dados de treinamento podem conter traços históricos de discriminação sistêmica intencional, decisões tendenciosas devido a diferenças injustas entre grupos e discriminação não intencional, ou podem ser amostras de dados que não representam todo o conjunto desejado. “Ao treinar as máquinas para reproduzirem as tarefas e formas humanas de reação e interação, corremos o risco de reforçar e reproduzir os estereótipos e preconceitos também se não houver esse cuidado para que os sistemas recebam valores humanos com senso de diversidade e inclusão”, aponta.

Rodrigo Kramper, líder da prática de advanced data & analytics solutions da ICTS Protiviti, acrescenta que os vieses existem, geralmente, em função das pessoas que podem trabalhar com dados sem qualidade adequada, sem objetividade ou em quantidade/ proporção inadequada. “Logo, quando utilizamos dados imprecisos ou incompletos, temos uma exposição maior a errar nas respostas dos algoritmos. Aqui, com certeza, vale a máxima de TI do “garbage in, garbage out”, ou seja, dados ruins de entrada geram resultados ruins de modelos”, diz.

Além da qualidade dos dados, Kramper aponta para o fator humano. “Todos temos uma visão de mundo e um erro comum é tentar interpretar os dados segundo a nossa visão e a nossa expectativa; e não pelo que eles representam”, reconhece Kramper.

O que fazer?

Muito dos problemas de viés advêm de distorções ou inadequações na obtenção dos dados e de sensibilidade para analisar os resultados dos algoritmos. Evitar o viés não é tarefa simples, porque nem sempre isso é obvio e perceptível durante o desenvolvimento, mas alguns cuidados ajudam a mitigar esses riscos. O primeiro é pensar em termos de proporcionalidade entre classes. “Se, por exemplo, estou desenvolvendo um algoritmo de reconhecimento facial e, se tenho um predomínio de imagens de homens, provavelmente o algoritmo terá dificuldades em reconhecer mulheres. O

mesmo aconteceria com etnias. A escolha das variáveis que compõem o modelo merece atenção", ensina Kramper.

Nas companhias, o viés deve ser endereçado primeiro com transparência, adverte Kramper, da ICTS Protiviti. "Equipes de ciências de dados têm de garantir a "explicabilidade" dos modelos, ou seja, mais e mais pessoas dentro da organização e de fora dos times de ciências de dados devem entender o objetivo dos modelos, as decisões de negócios tomadas por trás deles, quais são os dados utilizados, quais são os resultados e como os modelos chegaram nesse resultado", relata. Isso está longe de ser uma tarefa fácil, mas Kramper também reforça a necessidade de uma abordagem de governança de modelos. "Da mesma forma, precisamos ter mais clareza em relação aos impactos que o uso do modelo implicará não só no negócio, mas na sociedade e ter consciência das eventuais limitações impostas pelos modelos matemáticos", completa.

Expandir a base de dados de modo que fique o mais diversa possível é um passo importante para mitigar o viés. Por exemplo, em uma solução de recrutamento e seleção, é necessário que os dados que sejam representativos, com candidatos de todos os tipos de gênero, cor, condições socioeconômicas etc. "Deve-se observar também no mercado de inteligência de negócios. Ele acrescenta que as empresas, ao adotar Inteligência Artificial, deveriam colocar no ciclo operacional analítico a velocidade de produção e processos de governança necessários para garantir que não haja subjugamento.

"Hoje, para gerar dados em informação inteligente e colocar ela na ponta para ser consumida, você passa pelo ciclo de produção operacional analítico. Ou seja, vai construir modelo, pensar em colocar em produção e, depois que coloca em produção, a API começa a tomar decisões. Mas você tem a necessidade de monitorar o modelo para o ecossistema ficar vivo e eficiente", explica Maia. Entre os problemas a serem enfrentados que nem todo modelo, nem todo algoritmo de IA, vai dar um desempenho adequado para o problema, então, tem de escolher o algoritmo adequado à tarefa e treinar o modelo de forma adequada. Ou seja, tem de ser adequado e estar devidamente treinado. A escolha e o treinamento são importantes para mitigar o enviesamento", aponta Leandro Silva, do Mackenzie, explicando que é possível ter um algoritmo que seja bom, mas não estar bem preparado para resolver o problema. Para ele, as empresas estão olhando viés e ética de forma mais técnica e buscando meios de automatizar e melhorar, porque o viés é algo indesejado.

Bruno Maia é diretor de inovação da SAS, líder em software e serviços de análise de negócios e o maior fornecedor independente está o tempo que o processo leva, porque o dado, quando criado, já está sendo deteriorado. O quão mais rapidamente que puder ser e o quão mais cuidadoso e zeloso for será melhor para o modelo e mais precisa será a tomada de decisão.

Os desafios, segundo Maia, podem ser divididos em três grandes blocos: o primeiro é a questão do dado, abordando o contexto e qual bloco de significância de informação que está fazendo tomar decisão mais adiante; o segundo é a normativa, qual é a regra que se seguirá; e o terceiro é a rastreabilidade. "Se eu não cuido da base, da rastreabilidade da norma e da rastreabilidade do contexto, não consigo nem entender se estou sendo permissivo na questão de bias", enfatiza Bruno Maia.

Governança dos dados, portanto, mostra-se fundamental para se conseguir voltar no processo e interpretar o machine learning para saber os caminhos que ele fez, porque aquele comportamento automático aconteceu, uma vez que a máquina não tem o sentimento e nem o contexto social que os humanos têm.

“Tem de ficar clara a forma como a decisão é tomada. Se você usar processo de governança e colocar ética e moralidade como pilares principais de sua empresa, fica bem administrável e bem provável que mantenha bias longe”, diz Maia, reforçando que o monitoramento deve ser contínuo. “Devem-se detectar as falhas de bias, se possível, antes de entrar em produção, porque a máquina só toma decisão errada se você deixar, se você não rastrear. Então, tem de pensar até que ponto se pode reduzir a velocidade para obter processo mais seguro”, aponta.

Ao usar dados de baixa qualidade para treinar os algoritmos, o resultado será comprometido, porque, se os dados carregarem vieses, os modelos irão reproduzir os mesmos vieses. “O modelo é a representação fiel daquilo que o ser humano o criou. Daí, a importância da mitigação de viés desde a fase de preparação dos dados de treinamento utilizando metodologia e ferramentas para mitigação dos riscos. Precisamos continuar diversificando e ampliando o grupo de pesquisadores e cientistas que trabalham nessa tecnologia. É imperativo que a IA de hoje reflita os valores das populações para as quais foi criada para servir. Devemos continuar saindo de nossos silos, melhorando o desequilíbrio de gênero entre ciência da computação e IA e injetar um espírito de diversidade e inclusão na IA que a torne mais justa, precisa e transparente para todos os indivíduos”, aponta Fabrício Lira, da IBM Brasil.

A diversidade na construção e no treinamento dos modelos de IA é extremamente importante para diminuir os vieses. Mas, conforme ressalta Lira, não é apenas a quantidade que irá fazer a diferença. Ao falar de conjuntos de dados que serão fonte para o treinamento de algoritmos, o executivo da IBM aponta elementos essenciais para evitar o viés: qualidade dos dados; representatividade; volume adequado; e testes *(veja box)*. “Além disso, a boa prática envolve o uso de ferramentas, métodos e processos que ajudam a mitigar esse risco seja em tempo de desenvolvimento ou posteriormente em tempo de execução.”

À medida que a tecnologia evolui, novas formas de treinar a IA vêm sendo desenvolvidas com fins de facilitar a mitigação de vieses, bem como permitir maior facilidade de explicação dos modelos.

Atualmente, diz Lira, modelos complexos consomem enormes volumes de dados e poder computacional, o que inviabiliza uma séria de aplicações. “Temas como “small data”e“one-shotlearning”vêmevoluindo com velocidade. Isso irá trazer mais transparência e confiança para encorajar as empresas ao uso da IA em casos de uso mais complexos”, aponta o executivo da IBM.

Ademais, uma abordagem de curadoria de dados visa a garantir aspectos como relevância dos dados, precisão e atualidade dos dados e, por último, precisamos de avaliações periódicas de existência de viés. Isso se dá, por exemplo, com análises em relação a quais variáveis do modelo tem maior peso em seu resultado e validar se isso é adequado ou não, se faz sentido ou não, ou seja, precisamos alinhar a realidade com os dados usados para treinar o modelo.

Se os resultados do modelo não são aderentes à realidade ou não condizem com a realidade, dados mais recentes e adequados devem ser obtidos e utilizados.