Dados, estatísticas e algoritmos

Perspectivas e riscos da sua crescente utilização

Ana Frazão
Advogada. Professora de Direito Civil e Comercial da UnB. Ex-Conselheira do CADE.

O objetivo do presente artigo é propor uma reflexão sobre a crescente utilização de dados, estatísticas e algoritmos computacionais na compreensão e solução de problemas humanos, sociais, políticos, econômicos e jurídicos, a fim de se fazer um balanço entre as perspectivas e os benefícios que podem decorrer da referida alternativa, por um lado, e os riscos e perigos, por outro.

Se dados e estatísticas já apresentam significativa importância há bastante tempo, ambos têm agora nos algoritmos computacionais, como sequências de instruções computáveis para a solução de problemas, mais um importante instrumento de aplicação. Não é sem razão que os algoritmos computacionais têm sido utilizados para as mais distintas funções, que vão desde recrutamento de empregados e *matchmaking* em plataformas de encontros amorosos até complexas operações em mercados financeiros e decisões de condenações penais[1].

São várias as razões que justificam a sedução dos dados e dos números que nos é proposta pelas estatísticas e pelos algoritmos da economia digital. Em seu célebre livro *Thinking, Fast and Slow*, o papa da economia comportamental Daniel Kahnemann[2] mostra brilhantemente como, diante de todas as limitações da racionalidade humana e das suscetibilidades às ilusões e vieses cognitivos, a estatística pode ajudar a evitar o pensamento causal impróprio, que é a tendência das pessoas de aplicar o pensamento causal em situações que exigiriam o

[1] Ver: EVANS, David; SCHMALENSEE, Richard. *Matchmakers*: The new economics of multisided platforms. Boston: Harvard Business Review Press, 2016. Sobre a questão criminal, ver, dentre outros, https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/?mbid=social_fb_onsiteshare e http://www.bbc.com/portuguese/brasil-37677421.
[2] KAHNEMANN, Daniel. *Thinking, fast and slow* [edição eletrônica]. Nova Iorque: Farrar, Straus and Giroux, 2013.

raciocínio estatístico. Afinal, como as pessoas tendem a ver padrões onde eles não existem, a estatística seria uma ferramenta hábil a resolver esses problemas.

Daí a conclusão de Kahnemann de que diversos estudos têm mostrado que os tomadores de decisão humanos são inferiores a uma fórmula de previsão mesmo quando são informados sobre a pontuação sugerida pela fórmula. Importante razão da inferioridade do julgamento dos especialistas é que humanos são incorrigivelmente inconsistentes em fazer julgamentos sumários de informação complexa.

Em sentido semelhante, Philip Tetlock[3], em seu *Expert Political Judgement,* mostra a desconcertante conclusão de que pessoas que passam a vida estudando determinado assunto podem produzir previsões menos exatas do que macacos jogando dardos. Afirma o autor que, entre esses sujeitos e aqueles que distribuíram suas escolhas uniformemente entre as opções disponíveis, os grandes conhecedores de determinados assuntos fazem prognósticos apenas ligeiramente melhores. Uma das razões apontadas é que as pessoas que adquirem conhecimentos em determinada área desenvolvem uma "ilusão" acentuada de sua habilidade e se tornam "superconfiantes", além de terem maior dificuldade para admitir seus erros.

Em obra mais recente, *Superforecasting*, Tetlock e Gardner[4] ressaltam a preocupante utilização de modelos preditivos, entendendo que especialistas em economia e política costumam errar, em média, 85% de suas previsões de longo prazo, seja diante da dificuldade de se enxergar ao longe, seja porque analistas que tendem a organizar o seu raciocínio em torno de grandes ideias são normalmente guiados por pensamentos ideológicos, procurando enquadrar o problema em seus modelos predefinidos e tratando o resto dos dados como algo irrelevante.

A partir de tais conclusões, verificada a falibilidade da racionalidade humana para fazer julgamentos preditivos, tem-se um campo fértil para que julgamentos humanos sejam progressivamente substituídos por estatísticas e algoritmos computacionais. Vale notar que o próprio Kahnemann é um

---

[3] TETLOCK, Philip. *Expert political judgement*: How good is it? How can we know? Princeton: Princeton University Press, 2005.
[4] TETLOCK, Philip; GARDNER, Dan. *Superforecasting*: The art and Science of prediction. Nova Iorque: Crown: 2015.

entusiasta dessa possibilidade, ao afirmar que a hostilidade a algoritmos provavelmente irá abrandar à medida que seu papel na vida cotidiana continuar a se expandir.

Entretanto, o apreço às estatísticas e aos algoritmos computacionais vem acompanhado de uma série de ressalvas, muitas delas reconhecidas por Kahnemann[5], dentre os quais a confusão entre correlação e causalidade. Daí afirmar que as estatísticas produzem muitas observações que parecem pedir por explicações causais, mas que não se prestam a tais explicações, até porque muitos dos fatos do mundo devem-se ao acaso, incluindo acidentes de amostragem.

De forma convergente, Darrell Huff[6], em seu famoso *Como mentir com a estatística*, afirma: "A linguagem secreta da estatística, tão atraente em uma cultura voltada para fatos, é empregada para apelar, inflar, confundir e levar a simplificações exageradas. Métodos e termos estatísticos são necessários para relatar dados de tendências sociais e econômicas, condições de negócios, pesquisas de opinião e censos. No entanto, sem redatores que usem as palavras com honestidade e conhecimento, e sem leitores que saibam o que elas significam, o resultado só pode ser um absurdo semântico".

Daí o cuidado que se deve ter com o que Darrell Huff chama de "estatística corrompida", o que pode decorrer de inúmeras estratégias, tais como amostras pequenas ou com tendenciosidade embutida, médias bem escolhidas – que podem ocultar as desproporções entre os extremos da amostragem –, dentre inúmeras outras "técnicas". Daí as advertências do autor para a análise de alguns aspectos básicos de aferição da idoneidade da estatística, tais como quem a está apresentando, quais são as fontes, como a informação foi obtida, o que pode estar faltando na análise, dentre outros aspectos.

Todos esses cuidados devem se juntar ao esforço indispensável para se distinguir correlação de causalidade, uma vez que mesmo altas correlações podem não ter nenhum significado do ponto de vista causal. Aliás, sobre o assunto, vale consultar o site *Spurious Correlations*[7], que apresenta vários exemplos de correlações esdrúxulas, com altíssimos índices, como o de 99,79%,

---

[5] KAHNEMANN, Op. cit.
[6] HUFF, Darrell. *Como mentir com estatística*. Rio de Janeiro: Edições Financeiras S.A., 1968. p. 7.
[7] Disponível em: <http://tylervigen.com/spurious-correlations>

que corresponde à correlação entre gastos com pesquisa científica, espacial e tecnológica dos Estados Unidos e suicídios por enforcamento, estrangulamento e sufocamento.

Tal preocupação é especialmente relevante nos atuais tempos de "pós-verdades", em que tudo pode ser aceito facilmente, ainda mais se estiver ancorado em um número, cuja presença por si só já se mostra suficiente para aumentar a confiabilidade da informação.

Exemplo recente da experiência brasileira serve para alertar sobre os riscos da utilização de estatísticas sem os devidos cuidados. Trata-se da notícia da excessiva judicialização das relações de trabalho no Brasil, o que poderia ser comprovado pelo fato de termos 98% das ações trabalhistas do mundo, estatística que foi amplamente divulgada sem que tenham sido tomadas nenhuma das medidas de precaução sugeridas por Darrell Huff. Somente depois é que se percebeu que a estatística, além de não vir de fontes confiáveis, era totalmente incondizente com a realidade brasileira e mesmo com a realidade do mundo.

Entretanto, o episódio mostrou claramente não apenas que as pessoas podem confiar cegamente nas estatísticas, mesmo quando estas são manifestamente inidôneas, como também alertou para o risco da "absolutização do número", que é a tentativa de extrair conclusões apressadas e que não decorrem necessariamente da estatística apresentada, ainda que esta fosse correta.

Tais aspectos mostram claramente que as mesmas limitações da racionalidade que justificam a utilização das estatísticas nos assuntos humanos podem ser utilizadas para deturpar e corromper as estatísticas, da parte de quem as elabora e as difunde. Igualmente se pode verificar que as limitações da racionalidade humana podem fazer com que as estatísticas sejam indevidamente compreendidas pelos seus destinatários, tanto naquilo que pretendem demonstrar, como naquilo que muitas vezes procuram ocultar. Com efeito, não é raro que uma estatística seja utilizada para encobrir a realidade, destacando apenas um panorama parcial, normalmente a favor daquele a quem a estatística aproveita, a partir do qual se procura apresentar uma solução para o todo.

O problema potencializa-se com os algoritmos computacionais, que se baseiam em dados e correlações normalmente sigilosos e sem qualquer

transparência, motivo pelo qual podem utilizar dados incorretos ou falsos, bem como se prestar a reproduzir correlações que não correspondem a causalidades e, o que é mais grave, a reproduzir correlações que podem ser frutos de discriminações e uma série de injustiças da vida social.

Com efeito, em relação aos algoritmos computacionais, como bem aponta Harari[8] em seu *Homo Deus*, não se consegue nem mesmo responder à pergunta sobre sua origem e composição, mesmo porque, de maneira geral, constituem segredos de negócio. O autor cita o exemplo dos algoritmos do Google, que são desenvolvidos por equipes enormes, sendo que cada membro sabe apenas da sua parte do quebra-cabeça, mas não do todo. Ora, se nem mesmo quem está dentro é capaz de conhecer o algoritmo, com menor razão se pode cogitar de que os usuários possam ter algum acesso a tal tipo de informação ou mesmo capacidade técnica para compreendê-la.

Por outro lado, na medida em que são elaborados por homens, é inequívoco que a racionalidade limitada dos programadores pode transpor para as fórmulas dos algoritmos uma série de vieses e problemas cognitivos, os quais, diante da falta de transparência, não terão como ser objeto do devido escrutínio social, da crítica e do aprimoramento.

A falta de transparência é ainda mais reforçada quando se sabe que tais algoritmos são aperfeiçoados a partir da inteligência artificial, por meio da qual, com a aprendizagem automática e com as redes neurais artificiais, mais e mais algoritmos se desenvolvem independentemente, aprimorando a si mesmos e aprendendo com os próprios erros. Como bem resume Harari[9], "Eles [os algoritmos] analisam quantidades astronômicas de dados, que nenhum humano é capaz de abranger, e aprendem a reconhecer padrões e adotar estratégias que escapam à mente humana. O algoritmo-semente pode de início ser desenvolvido por humanos, mas ele cresce, segue o próprio caminho e vai aonde humanos nunca foram antes – até onde nenhum humano pode segui-lo".

Tal situação é inequivocamente geradora de riscos e perplexidades. Imagine-se um algoritmo desenvolvido para o recrutamento de pessoal em que

---

[8] HARARI, Yuval Noah. *Homo Deus*: uma breve história do amanhã [edição eletrônica]. São Paulo: Companhia das Letras, 2016.
[9] HARARI, Op. cit.

os perfis ideais dos candidatos foram convertidos em fórmula a partir de uma grande base de dados. Não seria nenhuma surpresa que o algoritmo desse maior peso a homens brancos para altos cargos, pois são eles que, de fato, ainda ocupam a maior parte das melhores posições. Não seria surpresa igualmente que, mantendo-se os referidos padrões sociais, os mecanismos de inteligência artificial atribuíssem uma crescente importância a tais aspectos no recrutamento. O grande problema de tal correlação é que ela obviamente não indica que homens brancos são melhores do que homens negros ou mulheres, mas reflete na verdade o resultado de aspectos culturais muito mais complexos, tais como a discriminação de raça e de gênero no mercado de trabalho.

Daí o fundado receio de que dados e correlações manejados por algoritmos possam estar sendo utilizados como veículos de manutenção de discriminações e injustiças, preservando os padrões do passado – ainda que equivocados – ao mesmo tempo em que comprometem as possibilidades do futuro em termos de desenvolvimento e emancipação social. E, o que é pior, na ausência de transparência quanto aos dados, critérios e correlações utilizados, os resultados práticos da aplicação de tais algoritmos computacionais podem ser insuscetíveis de um devido controle por parte do direito.

Somente essa constatação já mostra que a utilização de algoritmos computacionais, nos mais diversos campos, precisa ser acompanhada da devida e necessária reflexão. Considerando a magnitude do potencial dos algoritmos computacionais e mesmo a possibilidade de sua capacidade se sobrepor à dos seres humanos, Harari[10] propõe a seguinte pergunta: o que vai acontecer à sociedade, aos políticos e à vida cotidiana quando algoritmos não conscientes, mas altamente inteligentes, nos conhecerem melhor do que nós mesmos?

É certo que as perguntas propostas por Harari são de extrema complexidade e não caberia a este modesto ensaio tentar respondê-las. Entretanto, suas perguntas servem de alerta para a necessidade de se adotar uma visão mais cautelosa a respeito dos dados e das estatísticas, a fim de evitar a sedução enganosa dos números. Especificamente em relação aos algoritmos computacionais, suas perguntas igualmente servem de alerta para que reflitamos sobre o que podemos fazer para evitar que o nosso futuro seja definido por eles,

[10] HARARI, Op. cit.

sem que tenhamos nem mesmo a possibilidade de conhecer e criticar os dados e correlações que os alimentam e seus processos de aprimoramento futuro a partir da inteligência artificial.