"DESENVIESAMENTO" COGNITIVO PELA INTELIGÊNCIA ARTIFICIAL: INDIFERENÇA, ESTÍMULO OU COMPULSORIEDADE IMPOSTA PELO ORDENAMENTO JURÍDICO?

DEBIASING THROUGH AI AND THE LAW: INDIFFERENCE, INCENTIVE OR MANDATORY USE?

LUDMILA JUNQUEIRA DUARTE OLIVEIRA[1]
BRUNELLO STANCIOLI[2]

RESUMO

Neste artigo os autores analisam se a inteligência artificial pode auxiliar a afastar vieses cognitivos já identificados na tomada de decisões humanas, tornando-as mais acuradas e lógicas. Indagam ainda se, caso a inteligência artificial venha a auxiliar na tomada de melhores decisões, qual será a reação do ordenamento jurídico, que poderá variar da indiferença, passando pelo incentivo à utilização da inteligência artificial em casos de demonstrado incremento no processo decisório, chegando até mesmo a tornar obrigatório, via edição de leis, o uso em alguns casos. O artigo foi desenvolvido a partir da leitura crítica de textos sobre o desenvolvimento da inteligência artificial e suas repercussões na tomada de decisões humanas, culminando com uma projeção das implicações de tal uso para o Direito. Concluiu-se que a reação acerca da utilização da inteligência artificial nos processos decisórios com impactos jurídicos possivelmente dependerá da natureza dos interesses envolvidos e dos princípios característicos de cada área do Direito.

Palavras-chave: inteligência artificial; vieses cognitivos; desenviesamento.

ABSTRACT

This paper seeks to analyze if the artificial intelligence may contribute in debiasing human decision making, helping to make it more logical and precise. The authors will also verify if artificial intelligence can have a role in debiasing, and, in that case, what would be the Law' stand, whether it will be indifferent, incentive its use or impose it through regulation. The article was developed through the critical analysis of readings about AI and its repercussions on decision making, culminating with a projection of possible implications for the Law. The authors conclude that the use of AI in decision making with juridical repercussions will depend on the interests involved and the peculiar principles of each Law field.

Keywords: *debiasing; artificial intelligence; law.*

1 Mestre em Direito pela Universidade Federal de Minas Gerais (UFMG). Procuradora da República. ORCID iD: http://orcid.org/0000-0002-4251-0629.

2 Mestre e Doutor em Direito. Academic Visitor, Oxford University (2011-2012). Professor de Direito na Universidade Federal de Minas Gerais (UFMG).

Como citar esse artigo:/*How to cite this article*:
OLIVEIRA, Ludmila Junqueira Duarte; STANCIOLI, Brunello. "Desenviesamento" cognitivo pela inteligência artificial: indiferença, estímulo ou compulsoriedade imposta pelo ordenamento jurídico?. **Revista Meritum**, Belo Horizonte, v. 18, n. 2, p. 34-53, 2023. DOI: https://doi.org/10.46560/meritum.v18i2.7718.

1. INTRODUÇÃO

Neste artigo será analisado se a inteligência artificial poderá auxiliar a afastar *vieses cognitivos* já identificados na tomada de decisões humanas, tornando-as mais acuradas e lógicas. Os autores indagaram ainda se, caso a inteligência artificial venha a auxiliar na tomada de melhores decisões, qual será a reação do ordenamento jurídico, que poderá variar da indiferença, passando pelo incentivo à utilização da inteligência artificial em casos de demonstrado incremento no processo decisório, chegando até mesmo a tornar obrigatório, via edição de leis, o uso em alguns casos.

Desde a década de 1970, já foram produzidos vários estudos de psicologia, economia comportamental e ciência cognitiva que revelam que o processo decisório humano não é puramente lógico e racional, mas se vale de mecanismos heurísticos que facilitam as decisões. Esses mecanismos podem, no entanto, levar a desvios sistemáticos na formação de juízos, denominados vieses cognitivos[3].

Paralelamente à identificação dos vieses cognitivos, vem se desenvolvendo a inteligência artificial, cuja definição permanece nebulosa.

Para os fins do presente trabalho, será adotado um conceito mais amplo de inteligência artificial, abarcando tanto programas de computador quanto aplicações corporificadas em máquinas, capazes de agir de forma racional.

O surgimento da inteligência artificial impactou vários aspectos da vida individual e em sociedade, produzindo ganhos econômicos, gerando novos comportamentos e expectativas. É fato notório que a inteligência artificial é capaz de realizar cálculos e algoritmos de forma muito mais rápida e precisa que o cérebro humano. Contudo, em especial nos últimos anos, surgiram evidências de que também a inteligência artificial produz decisões enviesadas, ou seja, destoantes da expectativa racional de um processo decisório que considera determinados parâmetros predefinidos.

Por outro lado, em razão de sua capacidade analítica, alta velocidade e precisão na realização de cálculos matemáticos, a inteligência artificial é imune ao desenvolvimento de vários vieses cognitivos já identificados no comportamento humano. Neste contexto, serão analisadas hipóteses em que a inteligência artificial poderá auxiliar na tomada de decisão humana de modo a afastar a incidência de vieses cognitivos já identificados, processo conhecido em inglês como *debiasing*.

Por fim, indaga-se qual será a reação do ordenamento jurídico perante situações em que a inteligência artificial comprovadamente tornar a tomada de decisões humanas mais lógica e refratária à incidência de vieses cognitivos. Variando da completa indiferença, passando pelo estímulo à utilização de ferramentas de inteligência artificial na tomada de decisões humanas e chegando à imposição do uso por via legislativa, vislumbra-se que a inteligência artificial poderá contribuir em várias áreas do direito para melhorar o processo decisório humano, tornando-o mais lógico, racional e aderente aos valores do indivíduo.

O artigo foi desenvolvido a partir da leitura crítica de textos sobre o desenvolvimento e o funcionamento da inteligência artificial e suas repercussões na tomada de decisões humanas,

3 Haselton; Nettle; Andrews, 2005; Kahneman; Tversky, 1974; Kahneman; Tversky, 2012.

culminando com uma projeção das implicações de tal uso para o Direito. Assim, o artigo pode ser inserido no campo ainda em formação, que vem sendo denominado por alguns autores de "Neurodireito", consistente na disciplina transversal que terá o objetivo de reunir estudos das Ciências Cognitivas para subsidiar a revisão dos fundamentos e da dinâmica dos institutos jurídicos (Stancioli; Oliveira, 2020, p. 3). O escopo seria aprimorar as técnicas jurídicas com o que se sabe sobre o comportamento humano e oferecer respostas mais efetivas, desenhadas a partir das evidências coligidas e não com base em modelos ideais (Marden; Wykrota, 2018; Oliveira; Cardoso, 2018).

2. RACIONALIDADE E VIÉS COGNITIVO. INTELIGÊNCIA HUMANA E INTELIGÊNCIA ARTIFICIAL

Boécio forneceu célebre definição de pessoa como "*substância individual de natureza racional"*, que predominou durou toda a Idade média e ainda hoje reverbera (Boécio, 2005, p. 168). A natureza racional permeia o próprio conceito de ser humano, mas não há definição unívoca do que se entende por "racional" ou "inteligente", e muitas vezes cai-se numa tautologia que pouco contribui para a compreensão (Tacca; Rocha, 2018, p. 60[4]).

Não obstante a ausência de unicidade, o conceito etéreo do ser racional permeou o desenvolvimento das ciências, inclusive sociais. Na economia ganhou as vestes de *homo economicus*, enquanto no direito toma a forma do "homem médio". Na economia, a exigência de racionalidade do *homo economicus* é ainda mais aguda, parece se tratar de um ser com inteligência similar à de Albert Einstein, com a memória maior que a do computador *Big Blue* da IBM e com a força de vontade de Mahatma Gandhi (Thaler; Sunstein, 2008, p. 09-10). Sabe-se, contudo, que nem o homem médio nem o *homo economicus* podem ser encontrados desfrutando uma cerveja gelada no bar da esquina, em especial no final do mês, em um país em crise econômica…

Em que pese o conceito etéreo de ser racional tenha permeado o desenvolvimento científico, algumas áreas do conhecimento sempre se voltaram para as idiossincrasias humanas. Neste contexto, a partir da década de 1970, psicólogos, cientistas cognitivos e economistas comportamentais começaram a estudar o processo humano de decisão, jogando luz sobre como as informações necessárias são avaliadas e como juízos são formados. Desta forma, revelaram que muitas vezes seres humanos não usam a lógica e a razão, mais outros princípios heurísticos que simplificam a tomada de decisão (Kahneman; Tversky, 1974, p. 1124)[5]. Em geral, esses processos heurísticos são bastante úteis, pois reduzem a complexidade das tarefas envolvidas na tomada de decisão humana, mas muitas vezes podem levar a erros graves e sistemáticos, que configuram desvios-padrão hoje cognominados *vieses cognitivos* (*cognitive biases)*.

4 Neste sentido, veja-se a afirmação de Tacca e Rocha: "[…] podemos afirmar que o agente é de fato racional quando o seu desempenho passe a ser tão elevado quanto o de outro agente (humano) que possa executar aquela tarefa".

5 Os termos "princípios" e "regras" são usados de forma intercambiável neste tópico, no sentido lato de norma como um padrão, um parâmetro, e não com os significados antagônicos que os termos adquiriram no Direito. Na obra N*udge: improving decisions about health, wealth, and happiness,* os autores equiparam os princípios heurísticos a regras de experiência (r*ules of thumb)* utilizadas pelos seres humanos (2008, p. 10-11).

No texto "*Judgment under Uncertainty: Heuristics and Biases*", Kahneman e Tversky (1974) destacam que muitas decisões humanas são tomadas a partir de crenças na probabilidade de eventos incertos, cuja aferição é realizada com base em informações de validade limitada, processadas segundo regras heurísticas. Em um exemplo simples, os autores lembram que a distância de um objeto é parcialmente aferida pela clareza com a qual é visto. Quanto maior a acuidade com que se vê o objeto, mais próximo parece estar. Conquanto útil, a regra leva a erros sistemáticos, pois se as condições de visibilidade são ruins – por exemplo, em razão do mau tempo –, as distâncias são superestimadas, enquanto são subestimadas quando a visibilidade é boa, o que gera enviesamento no julgamento de distâncias a olho nu. Da mesma forma, a utilização de regras heurísticas para aferir probabilidades é útil na tomada de decisões humanas que exigem dados dessa natureza, mas também pode gerar enviesamento dos resultados.

Ao longo do citado artigo, Kahneman e Tversky (1974) abordam três processos heurísticos empregados pelo cérebro humano para aferir probabilidades (representatividade, disponibilidade, ancoragem e ajustamento) e os vieses que podem gerar, que serão sucintamente lembrados a seguir.

2.1 REPRESENTATIVIDADE

Para verificar a probabilidade de um fato em relação a outro, o cérebro humano afere as semelhanças entre eles, ou seja, verifica se um fato é representativo do estereótipo que se tem do outro. O exemplo de Thaler e Sunstein (2008, p. 17) ilustra o mecanismo: julgamos ser *mais provável* que um afrodescendente de 1,90 m seja um jogador de basquete americano do que um judeu de um 1,60 m, pois há muito mais jogadores de basquete que são *afrodescendentes* e *altos* do que *judeus* e *baixos*. Contudo, esse processo pode levar a incorreções, pois a aferição da semelhança (ou representatividade) deixa de sopesar vários fatores que devem ser considerados no juízo de probabilidade.

Kahneman e Tversky identificam seis vieses cognitivos que a heurística da representatividade pode gerar: insensibilidade à prévia probabilidade dos resultados; desconsideração do tamanho da amostragem; incompreensão da aleatoriedade; negligência à previsibilidade; ilusão de validade e incompreensão do fenômeno da regressão à média (Kahneman; Tversky, 1974, p. 1124-1127).

2.1.1 INSENSIBILIDADE À PRÉVIA PROBABILIDADE DOS RESULTADOS (OU NEGLIGÊNCIA DA TAXA-BASE)

Ao empregar a heurística da representatividade para aferir probabilidades, as pessoas tendem a *desconsiderar* o que sabem sobre a prévia probabilidade daquele resultado, atendo-se apenas às semelhanças do fato com o padrão de comparação.

No exemplo dos autores, ao solicitar que, considerada uma lista de possibilidades (agricultor, vendedor, piloto de avião, bibliotecário ou médico), fossem atribuídas probabilidades à profissão de um homem descrito como tímido e introspectivo, sempre prestativo mas pouco sociável, organizado e detalhista, espera-se que a resposta leve em conta o fato de que há muito mais agricultores do que bibliotecários na população. Contudo, os resultados demonstram

que usualmente negligencia-se tal dado, observando apenas a *similaridade* de descrição com o estereótipo das profissões de agricultor e bibliotecário.

2.1.2 INCOMPREENSÃO DA ALEATORIEDADE

A incompreensão da aleatoriedade leva as pessoas a esperar que uma sequência gerada por um processo aleatório tenha as características essenciais daquele processo, ainda que a sequência seja pequena e, portanto, sujeita a grandes desvios do padrão. Por outro lado, a incompreensão dos fenômenos aleatórios gera uma tendência a identificar padrões onde estes não existem, como exemplifica a "falácia do apostador"[6].

Um exemplo desse enviesamento ocorreu quando a *Apple* lançou o *iPod* com a opção *shuffle* (misturar) para ouvir seleções de músicas de forma aleatória e muitos usuários reclamaram que as listas não eram *realmente* aleatórias, pois um mesmo artista era tocado várias vezes e algumas músicas quase não tocavam (que é exatamente o que se espera de uma lista aleatória!). Cedendo aos anseios dos consumidores, em 2005 a empresa lançou o *smart shuffle*, que permitia ao usuário controlar o quanto seu *shuffle* era aleatório. À época, Steve Jobs justificou que estavam permitindo que o *shuffle* fosse menos aleatório para que *parecesse* mais aleatório[7].

2.1.3 NEGLIGÊNCIA À PREVISIBILIDADE

Aqui, a regra da representatividade leva o ser humano a negligenciar considerações sobre a *possibilidade das previsões*, novamente se concentrando apenas em aferir as semelhanças entre o evento e o estereótipo.

Os autores citam como exemplo pesquisa em que, a partir da descrição da performance de alunos de pedagogia em uma atividade didática, foi solicitado a um grupo que avaliasse o desempenho dos alunos e a outro que fizesse uma previsão de como os alunos estariam dali a cinco anos. Os juízos realizados pelos dois grupos foram *idênticos*, ou seja, mesmo cientes de que o desempenho em *uma* atividade acadêmica fornece muito pouca informação sobre o futuro profissional de alguém, os participantes da pesquisa fizeram as previsões apenas com base neste dado (Kahneman; Tversky, 1974, p. 1126).

2.1.4 INCOMPREENSÃO DO FENÔMENO DA REGRESSÃO À MÉDIA

O fenômeno da regressão à média foi primeiramente documentado por Francis Galton (1822-1911), que percebeu que, quando duas variáveis têm a mesma distribuição, se a média de uma delas desvia do meio por X unidades, a média da outra variável se alterará menos que X,

6 A "falácia do apostador" leva o indivíduo a crer que, quanto mais se acumulam as apostas perdidas, maior é a chance de que na próxima rodada saia vitorioso, apesar de ciente de que cada rodada é independente da anterior (Haselton, M. G.; Nettle, D.; Andrews, P. W. T**he evolution of cognitive bias.** In D. M. Buss, The Handbook of Evolutionary Psychology: Hoboken. Nova Jérsei: John Wiley & Sons Inc., 2005, p. 727).

7 Tal informação pode ser encontrada em artigos da área de informática, como no disponível em https://www.businessinsider.in/careers/news/spotify-made-its-shuffle-feature-less-random-so-that-it-would-actually-feel-more-random-to-listeners-heres-why/articleshow/74657263.cms. Acesso em: 11 set. 2021.

promovendo assim um retorno à média. O fenômeno já foi observado, por exemplo, em relação à altura de pais e filhos e performance de estudantes em provas consecutivas.

Apesar de ser observável ao longo da vida, as pessoas não desenvolvem intuições adequadas sobre o fenômeno. Por isso, deixam de prever a ocorrência da regressão à média em ocasiões em que certamente acontecerá. Por outro lado, quando identificam a ocorrência da regressão à média, usualmente inventam explicações causais duvidosas para o fato.

Kahneman e Tversky citam o seguinte exemplo de incompreensão da regressão à média: em uma discussão sobre treinamento de voo, instrutores experientes notaram que elogiar um pouso excepcionalmente suave era tipicamente seguido por um pouso pior na próxima tentativa, enquanto uma dura crítica depois de um pouso ríspido era normalmente seguido de uma melhora na próxima tentativa. Com base nessas constatações, os instrutores concluíram que elogios verbais são *prejudiciais* à aprendizagem, ao tempo em que punições verbais seriam benéficas, contrariamente ao que é consenso em psicologia e pedagogia. Não se atentaram, contudo, que é usual em testes consecutivos que, depois de uma performance ruim, há uma melhora, e que um desempenho extraordinário é usualmente seguido por um pior, *independentemente* da ocorrência de algum comentário verbal do instrutor (Kahneman; Tversky, 1974, p. 1127).

2.2 DISPONIBILIDADE

Neste segundo princípio heurístico, Kahneman e Tversky (1974, p. 1127-1128) apontam que, em algumas situações, o cérebro humano afere a frequência ou probabilidade de um evento pela *facilidade* com que sua ocorrência é recordada. Assim, se as pessoas facilmente se lembram de exemplos relevantes, acreditarão que o evento ocorre com maior frequência do que se não recordarem com naturalidade dos fatos.

Assim como as demais regras heurísticas, a disponibilidade é útil, pois exemplos de fatos mais frequentes são normalmente recordados de forma mais rápida e melhor do que casos mais raros. Contudo, também está sujeita a gerar resultados enviesados, pois a disponibilidade de exemplos é afetada por outros fatores além da frequência e disponibilidade, como maior cobertura pela imprensa, conexão pessoal com o fato, proximidade do evento.

Kahneman e Tversky (1974, p. 1127-1128) identificam quatro vieses cognitivos a que a heurística da disponibilidade pode induzir: viés em razão da memorabilidade dos exemplos; viés em razão da eficácia da busca de exemplos; viés em razão da capacidade de imaginar (*imaginability*) e correlação ilusória.

2.2.1 VIÉS EM RAZÃO DA MEMORABILIDADE

Quando a frequência de um evento é julgada pela facilidade com que exemplos de sua ocorrência são lembrados, os casos recordados com maior naturalidade são considerados mais recorrentes do que aqueles que, embora aconteçam com a mesma frequência, não são tão facilmente rememorados. Neste contexto, eventos que tiveram maior divulgação são lembrados com mais facilidade, fatos recentes têm mais impacto no nosso comportamento do

que aqueles ocorridos há mais tempo e experiências pessoais são mais marcantes que relatos de terceiros, desta forma influenciando como julgamos a frequência de tais ocorrências.

A título de exemplo, pode-se citar a comparação entre o carro e o avião como meios de transporte. Como os acidentes de avião costumam ser muito mais divulgados pela mídia que os de carro, muitas pessoas acreditam que o automóvel é mais seguro do que o avião como meio de transporte, enquanto a probabilidade de acidentes com este é, na verdade, bem menor.

2.2.2 VIÉS EM RAZÃO DA CAPACIDADE DE IMAGINAR (IMAGINABILITY)

Para aferir a frequência de um evento, algumas vezes não temos exemplos em nossa memória, mas podemos imaginá-los de acordo com determinadas regras. Nessas situações, a frequência é tipicamente aferida segundo a facilidade com que os exemplos são *criados*.

Kahneman e Tversky (1974, p. 1128) apontam que a capacidade de imaginar tem um importante papel na avaliação de riscos pelas pessoas. Por exemplo, uma expedição poderá ser avaliada como muito arriscada se as possíveis adversidades forem retratadas de maneira muito viva, apesar de a facilidade com que desastres são imaginados não refletir a probabilidade de sua ocorrência. Na mesma linha, os riscos envolvendo um empreendimento podem ser subestimados se alguns dos perigos possíveis são difíceis de imaginar.

2.2.3 CORRELAÇÃO ILUSÓRIA

Trata-se de viés cognitivo identificado por Chapman e Chapman (1967, p. 193-204) referente à frequência com que dois eventos ocorrem em conjunto. Verificou-se que quando há forte vínculo associativo entre os eventos, as pessoas julgam que sua ocorrência conjunta é maior do que quando não há tal ligação.

2.3 ANCORAGEM E AJUSTAMENTO

Ao abordar essa última regra heurística, Kahneman e Tversky (1974, p. 1128) sugerem que, para fazer estimativas numéricas, as pessoas usualmente partem de um valor inicial que é ajustado para produzir o resultado solicitado. Contudo, o ajuste desse valor inicial normalmente se mostra insuficiente, seja quando tal ponto de partida foi sugerido pela própria formulação do questionamento ou determinado espontaneamente pelo indivíduo.

Thaler e Sunstein (2008, p. 11-12) citam o exemplo referente à estimativa da população de uma cidade, aqui transportado para o Brasil para facilitar a compreensão. Se perguntarmos a um belo-horizontino qual a população de Uberlândia/MG, ele pode partir de um dado conhecido, como a população da capital mineira, cerca de 2,5 milhões de habitantes. Sabendo que Uberlândia é uma cidade grande, mas não do tamanho de Belo Horizonte, poderia estimar que a cidade tem um terço da população da capital, cerca de 800 mil habitantes. Mas se indagarmos a um morador de São João del-Rei/MG, que sabe a população da sua cidade (em torno de 90 mil habitantes) e que Uberlândia é maior, talvez cerca de três vezes maior, a resposta seria por volta de 270 mil habitantes (Uberlândia possui aproximadamente 580 mil habitantes).

A grande diferença entre as respostas reside exatamente na diversidade dos pontos de partida e insuficiência dos ajustes, verificados em várias pesquisas.

2.4 OUTROS VIESES COGNITIVOS

Além dos vieses cognitivos identificados como desvios-padrão dos processos heurísticos de julgamento, revelados inicialmente por Kahneman e Tversky (1974), hoje a denominação é usada para abarcar também outras situações. Assim, a terminologia engloba casos em que a incorreção decorre das limitações do cérebro humano em processar, de forma totalmente racional, determinadas informações, bem como quando se trata de mecanismo de adaptação evolutiva que acarreta menores custos com os eventuais erros do que a solução não enviesada (teoria do gerenciamento de erros). Naquele grupo estão incluídos os já citados exemplos de vieses referentes à aferição de probabilidades e neste as ilusões positivas, como o otimismo irreal e a sobreconfiança (Haselton; Nettle; Andrews, 2005, p. 726).

O *otimismo irreal* e a *sobreconfiança* levam as pessoas a acreditarem que têm ou terão desempenho acima da média em determinada atividade, mesmo quando cientes de que as chances de sucesso são pequenas e de que há variações naturais entre a performance dos que desempenham a atividade. O fenômeno já foi observado em questionários aplicados antes do início de cursos de pós-graduação(a maior parte dos estudantes reputava que ficaria entre os 20% melhores da turma); em relação à autoavaliação de professores em universidades (94% acreditava ser melhor que a média) e de motoristas (90% reputou conduzir melhor que a média) (Thaler; Sunstein, 2008, p. 27).

Thaler e Sunstein (2008, p. 28) destacam que o otimismo irreal é comum nos seres humanos, mas em algumas situações pode levar-nos a confiar excessivamente em uma suposta imunidade a perigos e negligenciar a adoção de cuidados preventivos.

2.5 INTELIGÊNCIA HUMANA E INTELIGÊNCIA ARTIFICIAL

Os vieses cognitivos apresentados nos tópicos anteriores revelam que a inteligência humana não segue um padrão único, racional e lógico na solução de problemas. Segundo a psicologia evolutiva, muitos dos mecanismos hoje identificados como vieses cognitivos surgiram como respostas a limitações de tempo e habilidade no processamento de informações necessárias à sobrevivência e reprodução humanas. Por isso, usualmente tais mecanismos são acionados pelo cérebro humano quando há restrições temporais e de dados, bem como quando são reduzidas as motivações para produzir resultados mais acurados (Haselton; Nettle; Andrews, 2005, p. 728).

Neste contexto, verifica-se que, apesar de o processo decisório humano estar sujeito à incidência de vieses cognitivos, muitas vezes é viável afastar o resultado enviesado, mediante adoção de cuidados para possibilitar que a tomada de decisões seja mais bem informada, melhor motivada ou guiada por um caminho apto a produzir resultados mais racionais, lógicos e adequados aos objetivos e valores do agente.

Paralelamente aos estudos sobre vieses cognitivos na inteligência humana, vem se desenvolvendo a cognominada inteligência artificial, cuja definição permanece nebulosa (Tacca;

Rocha, 2018, p. 58). O primeiro a utilizar a nomenclatura foi o professor de matemática John McCarthy que, junto com outros pesquisadores (Marvin Minsky, Nathan Rochester, Claude Shannon), organizou uma conferência de verão sobre o assunto, realizada em *Dartmouth College* em Nova Hampshire, EUA, em 1956. A denominação visava diferenciar o incipiente campo de estudo da já estabelecida área da cibernética[8] e sugeria que cada aspecto do processo de aprendizagem ou qualquer outro atributo da inteligência poderia, em tese, ser descrito com tamanha precisão de modo a viabilizar a construção de uma máquina que poderia simulá-lo (Kaplan, 2016, p. 13-14).

Olhando retrospectivamente, a denominação – que poderia ter sido originariamente substituída por "processamento simbólico" ou "computação analítica" (Kaplan, 2016, p. 17) – foi fortuita em atrair grande atenção (da mídia, do público e, consequentemente, de financiadores) para a área, apesar de trazer infindáveis e desnecessárias discussões advindas da comparação entre inteligência humana e inteligência artificial.

Atualmente, nem mesmo há consenso sobre a definição do termo "inteligência". Apesar de alguns pontos de vista alarmistas sobre a possibilidade de a inteligência artificial adquirir dimensões incontroláveis e ameaçar a existência humana (pelo menos da forma como a conhecemos[9]), hoje é comum a visão de que a inteligência artificial oferece mecanismos de solução de problemas muitas vezes melhores do que aqueles desenvolvidos pela inteligência humana. Está longe, no entanto, de se assemelhar ao cérebro humano. Ou seja, trata-se de dois conceitos diferentes, tornando muito pouco frutífero o debate sobre suas semelhanças e diferenças.

Mas afinal, o que é inteligência artificial? Também aqui não há consenso sobre o termo. Várias formulações referem-se à máquina ou ao programa de computador capaz de se comportar de forma que seria considerada inteligente caso se tratasse de seres humanos. Russel e Norvig (2010, p. 1-2) coletaram *oito* definições de inteligência artificial, que se diferenciam por focar na forma de pensar em oposição ao comportamento e na similitude com o ser humano em contraste com um padrão de racionalidade. Os autores demonstram preferência pela definição de inteligência artificial como o agente (máquina ou programa de computador) capaz de *atuar de maneira racional*, ou seja, de modo a obter o melhor resultado possível segundo regras lógicas. Justificam apontando que a busca por criar um agente capaz de desenvolver um comportamento racional é mais viável cientificamente, exatamente por que o comportamento humano não pode ser considerado inteiramente *racional*, mas sim bem adaptado ao ambiente que habitamos (Russell; Norvig, 2010, p. 4-5).

Não é objetivo, aqui, aprofundar a reflexão neste artigo, mas neste sentido Mercier e Sperber (2017, p. 7-12, 203-274) sustentam que a razão humana não é um mecanismo lógico ou um sistema de regras destinado a expandir e melhorar nossos conhecimento e decisões, mas sim um conjunto de processos informais, oportunistas e ecléticos, cuja principal função é simplificar e esquematizar argumentos intuitivos, para uso em nossas interações sociais. Assim, faz todo sentido diferenciar a inteligência humana de um modelo de pensamento padrão que funcione segundo regras lógicas.

8 Ciência que tem por objeto o estudo comparativo dos sistemas e mecanismos de controle automático, regulação e comunicação nos seres vivos e nas máquinas.

9 Como em HARARI, Yuval Noah. H**omo Deus:** Uma breve história do amanhã. Tradução: Paulo Geiger. São Paulo: Companhia das Letras, 2016 e em BOSTROM, Nick. S**uperintelligence:** Paths, Dangers, Strategies. *N*ew York: Oxford University Press, Inc., 2014.

A comparação entre inteligência artificial e inteligência humana revela que, apesar da coincidência de termos, são conceitos muito diversos, pelo menos no atual estado da tecnologia. Enquanto a inteligência humana foi se desenvolvendo ao longo de milhares de anos para enfrentar problemas concretos e se adaptar ao habitat ocupado pelos seres humanos, a inteligência artificial hoje é primordialmente construída para resolução de questões específicas selecionadas pelos seus criadores[10]. O ponto de contato é que ambos são processos cognitivos para resolução de problemas, envolvendo a coleta e processamento de dados, seguida da produção de um resultado. Neste contexto, fica evidente que é pouco útil a comparação inteligência artificial e humana, sendo muito mais proveitoso reconhecer que a inteligência artificial possui potencialidades ainda desconhecidas e pode contribuir muito para a melhora da qualidade da vida humana em vários aspectos, inclusive na tomada de decisões.

Neste sentido, Tacca e Rocha (2018, p. 66) sugerem que a inteligência artificial desencadeará mudanças acerca de quais tarefas e como são realizadas pelas pessoas. Muitas atividades que hoje são realizadas manualmente passarão a ser feitas por sistemas inteligentes. Enquanto outras, inclusive no campo legal, permanecerão fora do alcance das máquinas.

3. INTELIGÊNCIA ARTIFICIAL E ENVIESAMENTO: BIASED OR DEBIASED, *THAT'S THE QUESTION*

Conforme apontado no tópico anterior, inteligência artificial e inteligência humana são conceitos bem diversos. Ambos podem ser descritos como processos cognitivos para solução de problemas. Porém, a estrutura, o modo de funcionamento e os objetivos de cada um são muito diferentes. Em relação à inteligência humana, já foram identificados processos cognitivos que sistematicamente levam a decisões enviesadas e que muitas vezes não apresentam a melhor solução para o agente.

Mas será possível afirmar que os resultados produzidos pela inteligência artificial são sempre perfeitamente racionais e lógicos e desta forma isentos dos vícios que afetam o comportamento humano, tais como os vieses cognitivos?

A experiência acumulada ao longo dos anos revela que, apesar de serem capazes de produzir resultados muito mais rápidos e precisos do que a inteligência humana, também os programas de computador estão sujeitos à ocorrência de vícios e à produção de efeitos discriminatórios indesejados. Em artigo publicado ainda em 1996, Friedman e Nissenbaum (1996, p. 332) conceituaram programa de computador discriminatório como aquele que *sistemática* e *injustamente* discrimina contra alguns indivíduos ou grupo de indivíduos em favor de outros, ou seja, nega-lhes uma oportunidade ou bem ou atribui-lhes um resultado indesejado por fundamentos desarrazoados ou inapropriados. O fenômeno é conhecido em inglês como *machine bias*.

10 Ilustrando esse ponto de vista, vejam-se os já citados artigo de HASELTON, M. G.; NETTLE, D.; ANDREWS, P. W. T**he evolution of cognitive bias.** In D. M. Buss, The Handbook of Evolutionary Psychology: Hoboken. Nova Jérsei: John Wiley & Sons Inc., 2005 e a obra de MERCIER, Hugo; SPERBER, Dan. T**he enigma of reason.** *C*ambridge: Harvard University Press, 2017.

No estudo realizado por Friedman e Nissenbaum (1996, p. 332-336), foram identificadas três categorias de programas de computador discriminatórios: discriminações preexistentes, que incorporam preconceitos sociais ou individuais; discriminações técnicas, decorrentes de restrições ou aspectos técnicos; e discriminações emergentes, provenientes do uso dos programas.

Como exemplo recente de situação em que identificada a discriminação pela máquina, já foi noticiado que um algoritmo responsável por julgar o concurso de beleza mundial *Beauty.AI* (do qual participaram cerca de 6.000 pessoas em mais de 100 países) e que deveria ter considerado critérios objetivos como simetria facial e a presença de rugas para definir o concorrente mais atraente, apresentou resultado discriminatório, com relevante maioria de pessoas brancas e apenas uma negra, dentre 44 vencedores (Levin, 2016).

A presença de discriminação racial promovida por inteligência artificial também já foi registrada na utilização do programa *Compas* (*Correctional Offender Management Profiling for Alternative Sanctions*) nos EUA, que se destina a prever a possibilidade de reincidência e assim subsidiar a concessão de benefícios criminais, como fiança e livramento condicional. A meta do algoritmo é utilizar critérios mais objetivos e consistentes do que aqueles adotados por um agente humano, cuja subjetividade é intrínseca. Apesar de a inteligência artificial ser muito acurada em prever a reincidência, apurou-se que pessoas negras têm o *dobro* de chance de serem classificadas *erroneamente* como passíveis de reincidir, enquanto indivíduos brancos que reincidiram foram indevidamente considerados como de baixo risco quase *duas vezes* mais que os negros (Angwin; Larson; Mattu; Kirchne, 2016).

Outro exemplo provém de programas de reconhecimento facial. Pesquisas recentes realizadas com programas da Microsoft, IBM e Megvii da China apontam que a acurácia é muito maior para identificar homens brancos (superior a 90%) e muito menor em relação a mulheres negras (65%) (Buolamwini; Gebru, 2018). Neste campo, pode-se recordar ainda do aplicativo de reconhecimento de imagens *Google Photos* que, quando foi lançado em 2015, foi responsável por enorme gafe, pela qual a empresa passou meses se desculpando, ao classificar a imagem de um homem negro como um gorila (Dougherty, 2015).

Os exemplos de discriminação promovidos pela inteligência artificial são muito incômodos. Não só porque a expectativa geral é de que os programas de computador sejam melhores do que os seres humanos naquilo que se propõem a fazer, mas também porque seu uso está tão disseminado na sociedade que a discriminação promovida pelas máquinas pode alcançar proporções enormes.

As potencialidades de interferência da inteligência artificial na vida humana são tão grandes que alguns autores mencionam que estamos saindo de uma era da *internet* para a "sociedade do algoritmo", que será organizada em torno da tomada de decisões econômicas e sociais por algoritmos, robôs e inteligência artificial (Balkin, 2017, p. 3-4).

Por isso, é essencial assegurar que os resultados produzidos pela inteligência artificial sejam não apenas mais rápidos e acurados do que aqueles decorrentes do cérebro humano, mas também submissos a outros valores adotados pela sociedade, como a isonomia e a ausência de discriminação baseada em fundamentos razoáveis. Neste sentido, Friedman e Nissenbaum (1996, p. 345-346) sugerem que, além da segurança e confiabilidade, deve ser incluído como critério de avaliação da qualidade dos programas de computador a ausência de discriminação nos resultados produzidos.

Por outro lado (muito em razão de sua capacidade analítica, alta velocidade e precisão na realização de cálculos matemáticos), a inteligência artificial é imune ao desenvolvimento de vários dos vieses cognitivos já identificados no comportamento humano. Veja-se o exemplo dos vieses abordados acima, decorrentes de três processos heurísticos empregados pelo cérebro humano para aferir probabilidades, quais sejam, representatividade, disponibilidade e ancoragem e ajustamento.

Ainda que de forma simplificada, pode-se dizer que os computadores operam pela transformação de fórmulas matemáticas em algoritmo, ou seja, em *"conjunto metódico de passos que pode ser usado na realização de cálculos, na resolução de problemas e na tomada de decisões"* (Harari, 2016, p. 91). Diversamente do cérebro humano, que é intrinsecamente plástico, mutável, a memória dos computadores armazena digitalmente informações (Nicolelis; Cicurel, 2015, p. 13-25, 52-58).

Assim, a própria estrutura e funcionamento dos computadores evita a ocorrência dos erros praticados pelos seres humanos na formação de juízos de probabilidade. Programas de computador realizam com facilidade cálculos probabilísticos e não têm dificuldade para acessar informações gravadas em sua memória. Em suma, programas de computador criados para aferir probabilidades não necessitam fazer uso de heurísticas da representatividade, disponibilidade e ancoragem e ajustamento e, portanto, não estão sujeitos a incidir nos vieses cognitivos que daí podem decorrer.

Mesmo a inteligência artificial com emprego de técnicas de aprendizado de máquina (*machine learning*)[11], que muitas vezes gera resultados cujo processo de formação recebe a pecha de inconclusivo, inescrutável e opaco (Mittelstadt; Allo; Taddeo; Wachter; Floridi, 2016, p. 4-5), não incide em erros semelhantes àqueles gerados pela incidência dos vieses cognitivos abordados no presente artigo.

Nessa linha, veja-se que os exemplos de discriminação pela máquina citados acima não se confundem com os erros gerados por vieses cognitivos humanos, ainda que algumas vezes sejam decorrentes de preconceitos humanos anteriores, que acabaram incorporados no programa (discriminações preexistentes, na classificação de Friedman e Nissenbaum).

Conclui-se, assim, que a inteligência artificial pode auxiliar a afastar alguns vieses cognitivos já identificados na tomada de decisões humanas, tornando-as mais acuradas e lógicas, processo conhecido em inglês como *debiasing*, que pode ser traduzido como "*desenviesamento*".

Como pontuado acima, afirmar que a inteligência artificial pode contribuir para evitar a incidência de alguns vieses cognitivos identificados no processo decisório humano não implica concluir que aquela é *melhor* do que a inteligência humana. A inteligência artificial é criada especificamente para produção de resultados pontuais buscados pelo programador. Já o cérebro humano é fruto de um processo evolutivo que, segundo vários estudiosos, gerou módulos cognitivos adaptados para resolução de problemas concretamente enfrentados pelos seres humanos, dentre os quais a razão.

Neste cenário, Mercier e Sperber (2017, p. 7-12, 203-274) sustentam que o que se denomina razão humana não é um mecanismo para a realização de análises lógicas e tomada de

11 As técnicas de aprendizado de máquina (m*achine learning)* são programadas para criar e modificar regras de classificação de grandes bases de dados e neste processo podem identificar padrões obscuros e modificar a própria estrutura de forma não planejada pelo criador.

REVISTA MERITUM • v.18 • n.2 • p. 34-53 • Mai. - Ago. 2023

decisões melhores e bem fundamentadas, mas sim um instrumento que possibilita a interação entre as pessoas, característica fundamental da espécie e que viabilizou alcançar sua condição de predominância no mundo que habita (Harari, 2016, p. 150-158). Segundo os autores, a razão é o que nos possibilita justificar aos outros nossos pensamentos e ações, bem como produzir argumentos para garantir a cooperação de terceiros (Mercier; Sperber, 2017, p. 7, 107-202). Sob essa ótica, a presença de vieses cognitivos no pensamento humano não é necessariamente uma falha, mas a consequência acidental de um mecanismo adaptado para a resolução de um problema concreto enfrentado em algum momento pela espécie humana ao longo do processo evolutivo.

Ainda que sob o ponto de vista evolutivo os vieses cognitivos humanos não sejam considerados propriamente deficiências, não deve ser recusada a possibilidade de melhoria do processo decisório humano com o auxílio da inteligência artificial, deste modo produzindo decisões mais lógicas e providas de informações adequadas.

No item seguinte, será analisada tal possibilidade de aperfeiçoamento da tomada de decisões humanas pela inteligência artificial em algumas áreas do direito.

4. "DESENVIESAMENTO" COGNITIVO PELA INTELIGÊNCIA ARTIFICIAL: INDIFERENÇA, ESTÍMULO OU COMPULSORIEDADE IMPOSTA PELO ORDENAMENTO JURÍDICO?

No tópico anterior defendemos a possibilidade de a inteligência artificial contribuir para tornar a tomada de decisões humanas mais lógica e refratária à incidência de vieses cognitivos. Nesta parte, abordaremos outra indagação: qual será a mais adequada reação do ordenamento jurídico perante situações em que a inteligência artificial comprovadamente melhorar o processo decisório humano?

O ordenamento jurídico visa, por meio da imposição de normas, alterar o comportamento humano em determinada direção, alinhada aos interesses e objetivos considerados desejados pela sociedade. Sob essa premissa, mecanismos que possibilitem que as decisões e comportamentos humanos sejam direcionados em certo sentido, podem ser ou tratados pelo ordenamento jurídico com completa indiferença ou estimulados ou até mesmo impostos por via legislativa, a depender, dentre outros fatores, dos objetivos plasmados na ordem jurídica.

Veja-se, por exemplo, o uso de programas eletrônicos para o preenchimento da declaração de rendimentos para cálculo do imposto de renda no Brasil. Em 1991, a Receita Federal disponibilizou, como *alternativa* ao preenchimento da declaração em formulários de papel, programa de computador que permitia que os dados fossem informados por via eletrônica e entregues em meio magnético (disquetes). O uso do programa de computador tratava-se de mera faculdade, que em 1997 foi *estimulada* pela possibilidade de envio da declaração pela internet, ou seja, sem a necessidade de comparecimento físico do contribuinte em uma das unidades da Receita Federal.

Com a disseminação dos computadores e do acesso à *internet,* em 2011 a Receita Federal tornou *obrigatório* o preenchimento e envio da declaração de imposto de renda via programa específico disponibilizado na rede mundial de computadores (Receita Federal, 2014).

Neste caso do uso de programas de computador para preenchimento e remessa da declaração de rendimentos para cálculo do imposto de renda, a reação da ordem jurídica partiu da mera disponibilização da ferramenta, passando pelo estímulo ao uso, até torná-la obrigatória, em razão das evidentes vantagens que proporciona para o processamento das informações e cálculo do tributo, bem como no combate à sonegação.

E em relação à utilização da inteligência artificial para "desenviesamento" de decisões humanas, é possível prever qual será a reação do ordenamento jurídico? Com escopo de contribuir para a resposta a essa indagação, a seguir serão analisados três exemplos em que se vislumbra a possibilidade de utilização da inteligência artificial como instrumento para afastar a incidência de vieses cognitivos.

O primeiro exemplo é inspirado na obra "*Nudge: improving decisions about health, wealth, and happiness",* de autoria de Thaler e Sunstein (2008). A partir da análise de alguns vieses cognitivos já identificados no processo decisório humano, dentre os quais aqueles descritos neste artigo, os autores propõem a adoção de políticas públicas que auxiliem indivíduos na tomada de decisões raras e difíceis, para as quais não se recebe uma resposta imediata que permita readequar a decisão tomada ou quando é complicado traduzir aspectos da situação em termos de fácil compreensão (Thaler; Sunstein, 2008, p. 01). Um dos campos considerados férteis para tais políticas públicas, cognominadas "paternalismo libertário" (Thaler; Sunstein, 2008, p. 6-9), é o do direito previdenciário.

Thaler e Sunstein demonstram que, em geral, as pessoas têm enorme dificuldade em investir para garantir a formação de poupança adequada para a aposentadoria, problema que também é evidente no Brasil. Sugerem, assim, a adoção de políticas públicas que incentivem a adesão a planos de previdência, bem como que estimulem escolhas de investimento que garantam a formação de poupança adequada para a aposentadoria, a exemplo de programas de aumento progressivo de contribuições previdenciárias (Thaler; Sunstein, 2008, p. 1-21).

Com apoio nas considerações de Thaler e Sunstein, aventamos a possibilidade de adoção da inteligência artificial para indicar investimentos para previdência privada especificamente adequados ao usuário, a partir do exame de seu estilo de vida, rendimentos, condições de saúde e expectativa de vida. A inteligência artificial já tem condições de captar com facilidade dados sobre tais aspectos da vida do indivíduo, bem como pode ser programada a sopesar matematicamente esses fatores, para então indicar quais investimentos, em cada momento da vida, seriam os mais adequados para a formação de poupança para aposentadoria. Deste modo, a inteligência artificial pode contribuir para afastar os vieses cognitivos que afetam decisões desse tipo, em especial aqueles ligados aos juízos de probabilidade, otimismo irreal e a sobreconfiança.

O segundo exemplo colhe inspiração no artigo de Sunstein e Jolls, "*Debiasing through Law"*, em que os autores analisam hipóteses de "desenviesamento" pela lei, em que a *legislação* é utilizada para direcionar os indivíduos para a tomada de decisões mais lógicas e desprovidas de vieses cognitivos. Dentre as situações aventadas, os autores citam a utilização de normas no contexto de *segurança do consumidor*, de modo a obrigar que as informações sobre pro-

dutos perigosos sejam repassadas de forma contextualizada, com exemplos baseados em casos reais de acidentes e danos causados pelo uso inadequado do produto, invés de uma advertência genérica (Sunstein; Jolls, 2004, p. 207-216).

Partindo desta possibilidade, sugerimos o uso da inteligência artificial para criar alertas de segurança com conteúdo relevante e específico para o usuário, desta maneira tornando o teor da advertência mais memorável e disponível para lembrança pelo consumidor, em contraposição a avisos genéricos que muitas vezes são acintosamente ignorados. Com isso, poderiam ser afastados os vieses de otimismo irreal, sobreconfiança e relativos a juízos de probabilidade.

Os alertas poderiam utilizar dados pessoais, captados com a autorização do usuário, tanto no conteúdo da advertência quanto para aferir o *momento* em que seriam de fato necessários (por exemplo, quando o usuário demonstra alguma evidência de que fez ou pretende fazer uso inadequado do produto). A criação de alertas de segurança personalizados para o usuário, inclusive com aferição individualizada do momento e frequência da divulgação, poderia auxiliar a evitar o risco de banalização dos alertas, que contribuem para sua desconsideração pelo usuário. Outrossim, acreditamos que a definição do momento e frequência da divulgação do alerta de segurança pela inteligência artificial poderia contrabalancear o perigo de efeito reflexo sobre indivíduos que não incidem nos vieses de otimismo irreal, sobreconfiança e relativos a juízos de probabilidade e que, em razão das advertências personalizadas, poderiam adquirir a tendência a um pessimismo excessivo (Sunstein; Jolls, 2004, p. 228-231).

O terceiro exemplo situa-se no campo do direito ambiental, mais especificamente no licenciamento ambiental. Aqui, cogita-se o uso da inteligência artificial para realização de exames da probabilidade de determinados eventos e para cálculo de riscos de seus possíveis efeitos, desta maneira afastando possíveis vieses cognitivos relacionados aos juízos de probabilidade e otimismo excessivo na avaliação de impactos ambientais.

Neste último exemplo, desconhecemos estudos específicos que respaldem a efetiva incidência de vieses cognitivos na avaliação de impactos ambientais. Não obstante, como a atividade muitas vezes envolve a análise da probabilidade de ocorrência de determinados eventos, é razoável supor que também aqui possam incidir os vieses cognitivos identificados em avaliações semelhantes. A inteligência artificial poderia ser utilizada para melhorar o cálculo da probabilidade de determinados eventos e suas consequências, inclusive mediante uso de dados de empreendimentos semelhantes situados em outros locais. Aventa-se, outrossim, que a inteligência artificial possa contribuir na definição de medidas mitigadoras e compensatórias mais adequadas ao caso concreto.

Expostos esses três exemplos, retorna-se ao questionamento inicial: é possível prever qual será a reação do ordenamento jurídico em relação à utilização da inteligência artificial para "desenviesamento" de decisões humanas?

Considerando a diversidade de interesses envolvidos nos exemplos citados, não se vislumbra uma reação única pela ordem jurídica. Assim, no caso dos direitos previdenciário e do consumidor, em que estão envolvidos direitos individuais e tendo em conta que um dos objetivos da ordem jurídica é a proteção da autonomia, é mais adequado que o uso da inteligência artificial seja *estimulado*, mas *não imposto por lei*.

Ainda recordando as possibilidades acima expostas, o estímulo, no caso da inteligência artificial usada para assessoria de investimentos na poupança para aposentadoria, poderia

ser feito mediante concessão de benefícios tributários, por exemplo, a redução da alíquota de contribuição previdenciária ou dedução da base de cálculo do imposto de renda, nos moldes em que é feito em relação às despesas com planos de previdência privada (art. 11 da Lei nº 9.532/1997, com redação dada pela Lei nº 10.887, de 2004).

No caso do direito do consumidor, pode se pensar no estímulo via Poder Judiciário, por meio de decisões que considerem ser culpa exclusiva do consumidor o dano causado por ter ignorado advertência de segurança produzida e oferecida de forma personalizada para o usuário.

Esse prognóstico na seara do direito privado está alinhado às conclusões que Sunstein e Jolls (2004, p. 228-234) expõem em seu artigo. A estratégia de *estimular* a adoção da inteligência artificial para "desenviesamento" de decisões humanas sem, contudo, impor seu uso, evita o risco de atingir indivíduos que não sofrem a incidência de tais vieses cognitivos. Ademais, o estímulo, comparativamente à imposição por via legislativa, resguarda mais amplamente a esfera de autonomia do cidadão que, mesmo ciente das consequências possíveis, optar pelo risco de tomar decisões enviesadas.

Por outro lado, no caso do licenciamento ambiental, os interesses difusos envolvidos e os próprios princípios de direito ambiental, em especial os princípios da precaução e da prevenção, recomendam que a utilização da inteligência artificial no "desenviesamento" da análise dos impactos e riscos envolvidos seja imposta pela ordem jurídica. Para evitar o incremento dos custos do licenciamento ambiental de forma indiscriminada, a exigência do uso da inteligência artificial poderia ser deixada a cargo dos próprios órgãos licenciadores, na forma de condicionante à concessão da licença ambiental. Desta forma, os órgãos licenciadores poderiam avaliar, sob a ótica dos princípios da precaução e da prevenção, quando de fato é recomendável a ferramenta, com especial atenção para empreendimentos comprovadamente mais arriscados como, por exemplo, a mineração em grande escala; a construção de barragens e hidrelétricas; grandes obras de engenharia civil.

Todavia, a adoção da inteligência artificial deve ser feita de forma cuidadosa, pois não se pode olvidar que é capaz de gerar resultados discriminatórios, como ilustram os exemplos citados na terceira parte deste artigo. Da mesma forma que a matemática, que se utiliza de uma linguagem que nem todos dominam, programas de computador (que devem muito de seu funcionamento à matemática) desfrutam do que Cathy O'Neil cognomina "autoridade do inescrutável", que se traduz na impossibilidade de se questionar aquilo que não se compreende (O´Neil, 2019).

Não obstante, especificamente no que se refere às possibilidades de "desenviesamento" do processo decisório humano, acreditamos que as vantagens na utilização da inteligência artificial superam os riscos de impactos negativos. Isso porque o debate ético sobre as consequências do uso da inteligência artificial já está na pauta da sociedade e das grandes empresas do setor de tecnologia da informação. Consoante explicitado ao longo do texto, desde a década de 1990 a discriminação pela máquina é objeto de estudo e cada vez mais o critério da ausência de discriminação nos resultados produzidos é incorporado na avaliação da qualidade dos programas de computador, como sugerido por Friedman e Nissenbaum (1996, p. 345-346).

Ademais, vários setores da sociedade têm se mostrado críticos e vigilantes quanto aos efeitos discriminatórios do uso da inteligência artificial. Há hoje crescente campo de estudo da "ética dos algoritmos" (Mittelstadt; Allo; Taddeo; Wachter; Floridi, 2016), em que autores como

REVISTA MERITUM • v.18 • n.2 • p. 34-53 • Mai. - Ago. 2023

O'Neil (2019) apontam para os riscos da utilização indiscriminada da inteligência artificial na tomada de decisões em áreas diversas. Conquanto analisar e trazer à tona a questão não necessariamente implique corrigir o resultado discriminatório gerado pelo uso do algoritmo, muitas vezes serviu para impulsionar a adoção de medidas para alterar o programa de computador e cessar o dano, como no caso do *Google Photos* e no recente episódio envolvendo o *Facebook* (Lapowsky; Matsakis, 2018).

Além disso, o resultado discriminatório produzido pela inteligência artificial pode ser mais facilmente corrigido, com a melhoria dos bancos de dados usados no treinamento, no caso do aprendizado de máquina, ou a reprogramação do algoritmo. Diversamente, o "desenviesamento" cognitivo do cérebro humano é bastante oneroso, pois exige concentração e foco dos indivíduos e a adoção de estratégias mentais na tomada de cada decisão. A mera ciência de que existe a possibilidade de incidirmos em vieses cognitivos não gera automaticamente o "desenviesamento" do raciocínio. Neste sentido, em seu multicitado artigo, Kahneman e Tversky (1974, p. 1125-1126) destacam que mesmo pesquisadores experientes, conhecedores das regras de probabilidade e dos desvios em que sistematicamente incidimos, demonstraram o viés de incompreensão da aleatoriedade, ao sustentar a expectativa de que uma hipótese sobre determinada população estaria estatisticamente bem representada *independentemente da amostragem*.

Já resultados discriminatórios gerados pela inteligência artificial podem ser inclusive prevenidos pelo programador e usuário. Neste contexto, Friedman e Nissenbaum (1996, p. 345-346) destacam que o debate ético e a adoção de critério referente à ausência de discriminação nos resultados produzidos para avaliação de algoritmos dão respaldo aos programadores para que possam se posicionar em busca da prevenção ou mesmo recusar-se a produzir programas que venham a gerar tais resultados indesejados.

Recentemente, reportagem do *The Guardian on line* divulgou artigo publicado por programadores, dentre os quais Moritz Hardt, pesquisador sênior do *Google*, que propõem um teste para detectar e uma solução para corrigir discriminação pela máquina em casos de aprendizado supervisionado[12].

Em suma, além de suas características intrínsecas impedirem o desenvolvimento de muitos dos vieses cognitivos identificados no comportamento humano, eventual enviesamento ou efeito discriminatório detectado em decisões tomadas pela inteligência artificial pode ser prevenido e, se verificado, corrigido com mais facilidade.

5. CONCLUSÃO

O principal objetivo neste artigo é destacar a possibilidade de a inteligência artificial contribuir em várias áreas do direito para afastar a incidência de vieses cognitivos e assim melhorar o processo decisório humano, tornando-o mais lógico e aderente aos valores e objetivos do indivíduo e da sociedade.

12 A íntegra do artigo E*quality of Opportunity in Supervised Learning* está disponível em https://arxiv.org/abs/1610.02413.

A partir de alguns exemplos da adoção da inteligência artificial para "desenviesamento" das decisões humanas, fez-se uma análise de qual seria a reação mais adequada do ordenamento jurídico em face dessas possibilidades. A conclusão a que se chega é de que é muito pouco provável que a ordem jurídica reaja de maneira única quanto ao uso da inteligência artificial no "desenviesamento" de decisões humanas em todas as áreas do direito. A opção entre a indiferença, o estímulo e a imposição do uso da inteligência artificial como instrumento para melhoria do processo decisório humano dependerá da natureza dos interesses envolvidos e dos princípios característicos de cada área.

Nessa linha, sugere-se ser mais adequado que o ordenamento jurídico apenas *estimule* a adoção da inteligência artificial no "desenviesamento" de decisões na seara do direito privado, invés de impor a utilização da ferramenta por via legislativa. Desta forma, acredita-se que será mais amplamente resguardada a autonomia individual, bem como evitado o risco de atingir pessoas que não sofrem a incidência de tais vieses cognitivos e poderiam ser oneradas ou prejudicadas com a interferência indiscriminada da inteligência artificial em seu processo decisório. Já na esfera do direito público, em que os interesses que se busca tutelar são difusos e preponderam sobre os privados, em algumas situações será apropriado impor a adoção da inteligência artificial para o incremento do processo decisório humano.

Veja-se que não estamos advogando a adoção indiscriminada da inteligência artificial em substituição à tomada de decisões pelos indivíduos, apesar de muitos autores já apontarem ser este um cenário inevitável[13]. O que se defende neste artigo é o *potencial* da inteligência artificial para contribuir na melhoria do processo decisório humano, o que exige que o resultado seja não apenas mais lógico e acurado do que o obtido pelo cérebro humano, mas também isento de efeito discriminatório.

Consoante explicitado acima, não se pode olvidar que o uso indiscriminado da inteligência artificial pode gerar resultados discriminatórios, que são potencializados pela utilização disseminada e crescente dos algoritmos na tomada de decisões em todas as áreas da sociedade. Não obstante, acreditamos que os riscos de consequências negativas pelo uso da inteligência artificial são contrabalanceados pelo fato de que suas características intrínsecas impedem o desenvolvimento de muitos dos vieses cognitivos identificados no comportamento humano, aliado à possibilidade de prevenção e correção, com mais facilidade, de eventual enviesamento ou efeito discriminatório detectado em decisões tomadas pela inteligência artificial.

Talvez esta posição seja ingenuamente otimista, ao acreditar ser possível mitigar os danos que inevitavelmente serão causados pelo uso disseminado da inteligência artificial na tomada de decisões humanas. Contudo, como aponta Balkin (2017, p. 1446), a invasão da inteligência artificial em todos os setores da vida humana, inclusive do direito, parece ser inevitável. Só nos resta identificar o que será possível fazer para evitar maiores danos e esperar pelo melhor.

13 Nesse sentido: BALKIN, Jack M., T**he Three Laws of Robotics in the Age of Big Data.** Ohio State Law Journal, v. 78, n. 592, 2017, p. 3-4; HARARI, Yuval Noah. H**omo Deus: Uma breve história do amanhã.** Tradução: Paulo Geiger. São Paulo: Companhia das Letras, 2016. p. 383-399.

REFERÊNCIAS

ANGWIN, Julia; LARSON, Jeff; MATTU, Surya; KIRCHNE, Lauren. Machine Bias. **ProPublica**, 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 23 jun. 2018.

BALKIN, Jack M. The Three Laws of Robotics in the Age of Big Data. **[Ohio]Ohio State Law Journal**, v. 78, n. 592, 2017.

BOSTROM, Nick. **Superintelligence**: Paths, Dangers, Strategies. New York: Oxford University Press, Inc., 2014.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. **Proceedings of Machine Learning Research**. v. 81, 2018. Disponível em: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Acesso em: 11 jul. 2021.

DEVLIN, Hannah. Discrimination by algorithm: scientists devise test to detect AI bias. **The Guardian**. Disponível em: https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias. Acesso em: 11 set. 2021.

DOUGHERTY, Conor. Google Photos Mistakenly Labels Black People 'Gorillas'. **The New York Times**. Bits. 1 jul. 2015. Disponível em: https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/. Acesso em: 11 jul. 2021.

FRIEDMAN, Batya; NISSENBAUM, Helen. Bias in Computer Systems. **Journal ACM Transctions on Information Systems (TOIS)**, v. 4, n. 03, p. 330-347, 1996.

HARARI, Yuval Noah. **Homo Deus**: Uma breve história do amanhã. Tradução Paulo Geiger. São Paulo: Companhia das Letras, 2016.

HASELTON, M. G.; NETTLE, D.; ANDREWS, P. W. The evolution of cognitive bias. *In*: D. M. Buss (ed.). **The Handbook of Evolutionary Psychology**: Hoboken. New Jersey: John Wiley & Sons Inc., 2005. p. 724–746.

LAPOWSKY, Issie; MATSAKIS, Louise. You Can Now See All the Ads Facebook Is Running Globally. **Wired**. Business. 28 jun. 2018. Disponível em: https://www.wired.com/story/facebook-aims-more-transparency-view-ads-feature/. Acesso em: 11 jul. 2021.

LEVIN, Sam. A beauty contest was judged by AI and the robots didn't like dark skin. **The Guardian**, San Francisco, 8 set. 2016. Disponível em: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. Acesso em: 11 jul. 2021.

KAHNEMAN, Daniel; TVERSKY, Amos. **Judgment under Uncertainty**: Heuristics and Biases. Science, New Series, v. 185, n. 4157, p. 1124-1131, 1974.

KAHNEMAN, Daniel; TVERSKY, Amos. **Rápido e devagar**: duas formas de pensar. Rio de Janeiro: Objetiva, 2012.

KAPLAN, Jerry. **Artificial Intelligence**: What Everyone Needs to Know. Oxford University Press, 2016.

KELLY, Kevin. **The Myth of a Superhuman AI**. Disponível em: https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/. Acesso em: 2 jul. 2018.

MARDEN, Carlos; WYKROTA, Leonardo Martins. Neurodireito: o início, o fim e o meio. **Revista Brasileira de Políticas Públicas**, Brasília, v. 8, n. 2, p. 48-63, 2018. Disponível em: ht-tps://www.publicacoesacademicas.uniceub.br/RBPP/issue/view/244. Acesso em: 16 dez. 2019.

MERCIER, Hugo; SPERBER, Dan. **The enigma of reason**. Cambridge: Harvard University Press, 2017.

MITTELSTADT, Brent Daniel; ALLO, Patrick; TADDEO, Mariarosaria; WACHTER, Sandra; FLORIDI, Luciano. **The ethics of algorithms**: Mapping the debate, 2016. Disponível em: https://doi.org/10.1177/2053951716679679. Acesso em: 2 jul. 2018.

NICOLELIS, Miguel A. L.; CICUREL, Ronald. **The Relativistic Brain**: How it Works and why it cannot be simulated by a Turing Machine. Montreux: Kios Press, 2015.

OLIVEIRA, Thaís de Bessa Gontijo de; CARDOSO, Renato César. Consiliência e a possibilidade do neurodireito: da desconfiança à reconciliação disciplinar. **Revista Brasileira de Políticas Públicas**, Brasília, v. 8, n. 2, p. 116-142, 2018. Disponível em: https://www.publicacoesacademicas.uniceub.br/RBPP/issue/view/244. Acesso em: 16 dez. 2019.

O'NEIL, Cathy. The Authority of the Inscrutable: An Interview with Cathy O'Neil. [Entrevista concedida a] Carlos Delclós. **Centre de Cultura Contemporània de Barcelona**, Barcelona, 22 jan. 2019. Disponível em: https://lab.cccb.org/en/the-authority-of-the-inscrutable-an-interview-with-cathy-oneil/. Acesso em: 11 jul. 2021.

RECEITA FEDERAL. **Cronologia do imposto de renda**, 2014. Disponível em: http://idg.receita.fazenda.gov.br/sobre/institucional/memoria/imposto-de-renda/cronologia-do-imposto-de-renda/arquivos-e-imagens/2014.jpg. Acesso em: 11 jul. 2021.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence**: A Modern Approach. 3. ed. New Jersey: Prentice-Hall, 2010.

SCHERER, Matthew U. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. **Harvard Journal of Law & Technology**, v. 29, n. 2, 2016. Disponível em: https://ssrn.com/abstract=2609777 ou http://dx.doi.org/10.2139/ssrn.2609777. Acesso em: 2 jul. 2018.

SCHNEPS, Leila; COLMES, Coralie. **Math on trial**: how numbers get used and abused in the courtroom. New York: Basic Books, 2013.

STANCIOLI, Brunello Souza. **Renúncia ao Exercício de Direitos de Personalidade**: ou Como Alguém se Torna o que Quiser. Belo Horizonte: Del Rey, 2010.

STANCIOLI, Brunello Souza; OLIVEIRA, Ludmila. **Neurodireito e negócios jurídicos**. Belo Horizonte: Arraes Editores, 2020.

SUNSTEIN, Cass R.; JOLLS, Christine. Debiasing through Law. **John M. Olin Program in Law and Economics Working Paper**, n. 225, 2004.

TACCA, Adriano; ROCHA, Leonel Severo. Inteligência Artificial: Reflexos no sistema do Direito. **Revista do Programa de Pós Graduação em Direito – NOMOS**, v. 38, n. 2, 2018. Disponível em: http://www.periodicos.ufc.br/nomos. Acesso em:

THALER, Richard H.; SUNSTEIN, Cass R. **Nudge**: improving decisions about health, wealth, and happiness. New Haven: Yale University Press, 2008. Arquivo Kobo.

Dados do processo editorial

▪ Recebido em: 10/03/2020

▪ Controle preliminar e verificação de plágio: 11/05/2020

▪ Avaliação 1: 08/06/2020

▪ Avaliação 2: 02/06/2021

▪ Decisão editorial preliminar: 04/06/2021

▪ Retorno rodada de correções: 13/09/2021

▪ Decisão editorial/aprovado: 01/05/2022

Equipe editorial envolvida

▪ Editor-chefe: 1 (SHZF)

▪ Editor-assistente: 1 (ASR)

▪ Revisores: 2