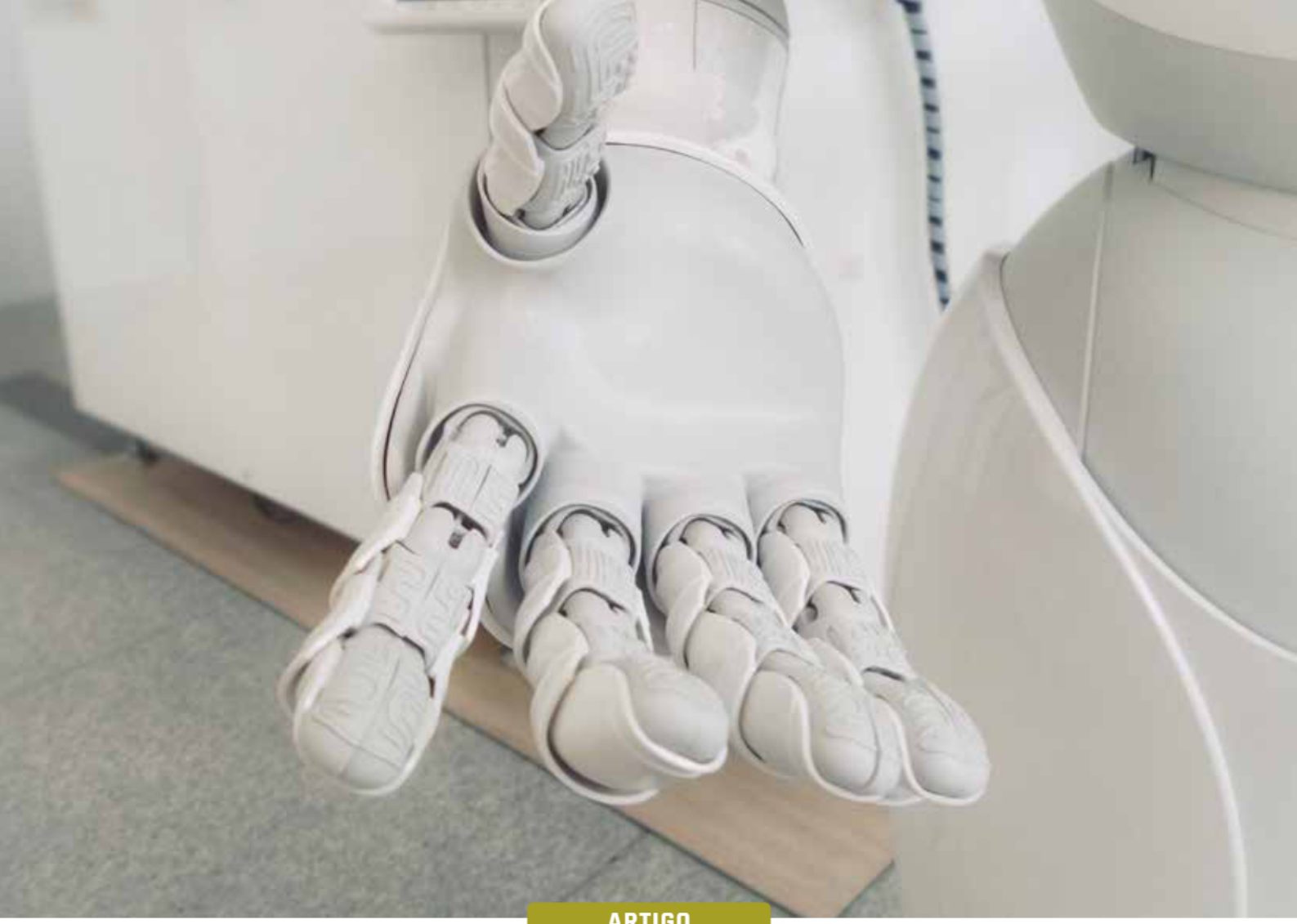

ARTIGO

# ÉTICA E INTELIGÊNCIA ARTIFICIAL

**POR**

Ana Cristina Bicharra Garcia

cristina.bicharra@uniriotec.br

A transformação digital vem fomentando o uso de técnicas de Inteligência Artificial (IA) por empresas e por governos. A realidade é que o cidadão mal se dá conta que interage com sistemas inteligentes o tempo todo, seja numa simples compra de cartão de crédito, seja recebendo dicas no seu canal preferido de *streaming*. O contexto da pandemia do COVID-19 impulsionou o uso de IA. No monitoramento de casos de infectados pelo vírus, governos utilizaram técnicas de reconhecimento facial e Big Data para controlar o fluxo de pessoas nas cidades e assim frear o contágio da doença [1]. Já médicos e cientistas se utilizaram do aprendizado de máquina para, a partir

de dados sobre a evolução dos casos encontrados, melhorar o diagnóstico e o prognóstico de novos infectados [2].

Entretanto, apesar dos avanços e benefícios que a IA, em especial o aprendizado de máquina, vêm trazendo, pesquisadores têm alertado para exemplos de vieses e preconceitos exacerbados por sistemas inteligentes. Neste artigo, vamos discutir o uso da Inteligência Artificial e os vieses sociais que podem estar contidos na enorme massa de dados utilizada pelos sistemas inteligentes e algoritmos de aprendizado de máquina.

## Como funciona a Inteligência Artificial e o aprendizado de máquinas

Para discutirmos o uso da IA, seus benefícios e suas limitações, devemos antes entender o seu funcionamento. A IA é uma área da computação voltada a desenvolver algoritmos e sistemas capazes de realizar tarefas que demandam habilidades associadas à inteligência humana. Dentre os exemplos mais conhecidos do uso da IA, encontramos a capacidade de poder se comunicar conosco na nossa linguagem, como os assistentes pessoais dos nossos celulares ou perceber e interpretar o mundo, como no reconhecimento de imagens realizado pelos carros autônomos. O emprego de técnicas de IA deve fazer com que a máquina possa ainda planejar sequências de atividades para alcançar metas, como nos sistemas inteligentes

Se a máquina receber **dados e informações** carregados de vieses e preconceitos de raça, de gênero, de escolha sexual, de forma física ou de qualquer outro traço, ela irá não só aprender com eles como perpetuá-los, durante o seu processo de aprendizado, quando exposta a novos dados.

que sabem jogar xadrez; raciocinar para resolver problemas complexos, como nos sistemas de diagnóstico médico; e, é claro conseguir aprender a fazer tudo isso sozinha.

A máquina será capaz de aprender se a ela for definido o passo a passo da tarefa, um algoritmo, assim como o ser humano aprende dos livros. Numa outra abordagem, a do aprendizado de máquina, em vez de modelar e ensinar o computador em cada etapa do processo, são fornecidas instruções de como aprender a partir de exemplos e dados. Isso significa que as máquinas podem ser usadas para tarefas novas e complicadas sem que seja programado manualmente o passo a passo de solução. Ela deve depreender do histórico de soluções qual o padrão do problema e qual deve ser o processo de solução.

Dado um grande conjunto de dados para o treinamento, um algoritmo de aprendizagem de máquina gera um modelo capaz de mapear entradas em

saídas. Este modelo matemático é composto por um sistema de equações não lineares com descontinuidade. O papel do algoritmo de aprendizagem é definir os valores ótimos para os coeficientes dessas equações de tal forma a bem reproduzir os exemplos do conjunto de treinamento, mas almejando a generalização. Parece um contrassenso, mas um modelo gerado que cubra 100% dos casos de treinamento é visto com reservas por se considerar que tal modelo pode ter apenas "decorado os exemplos" (*overfit*) sem ter extraído o padrão de mapeamento.

Os modelos de aprendizagem de máquina podem ser usados em tarefas e em domínios distintos. Um mesmo algoritmo de aprendizagem de máquina pode gerar modelos distintos dependendo da base de dados usada no treinamento do modelo, como por exemplo um sistema de diagnóstico de câncer para o domínio da saúde e um sistema de previsão do comportamento de ações para o mercado financeiro. Portanto, no aprendizado de máquina, os dados desempenham um papel fundamental: quanto mais dados (confiáveis) disponíveis para treinar o algoritmo, melhor será o modelo gerado por ele.

É justamente para este ponto que queremos chamar a atenção. A inteligência da máquina depende da qualidade dos dados e dos exemplos a que ela é submetida, e vai reproduzir o conhecimento que está impregnado nesses dados. Não é o suficiente se garantir que os dados estejam corretos. Esta seria a premissa básica, mas não é suficiente. Se a máquina receber dados e informações carregados de vieses e preconceitos de raça, de gênero, de escolha sexual, de forma física ou de qualquer outro traço, ela irá não só aprender com eles como perpetuá-los, durante o seu processo de aprendizado, quando exposta a novos dados. Nos exemplos a seguir, procuraremos apontar algum dos riscos de se aplicar indiscriminadamente a Inteligência Artificial sem discutir os dilemas éticos entremeados na enorme massa de dados que circulam pelos sistemas inteligentes do mundo todo.

### Dados não são neutros

Em 2016, um concurso de beleza chamou a atenção da mídia, pois se colocava como o primeiro certame cujos julgadores seriam máquinas, isto é, o júri seria composto por robôs. A novidade divulgada pelos realizadores era que o júri seria composto

exclusivamente por agentes artificiais (júri-robô) gerados por inteligência artificial. Os robôs tomariam decisões a partir de critérios objetivos e com isso tirar qualquer tipo de subjetividade na escolha. Esses robôs foram treinados para avaliar rugas, simetria facial, medidas faciais e uniformidade na coloração da pele antes de escolherem os homens e mulheres vencedores, considerando as várias categorias desde os 18 aos 69 anos de idade. Dessa forma, a promoção do evento garantia que o júri do projeto Beauty. AI[1] escolheria as concorrentes mais atraentes, sem preconceitos ou aspectos socioculturais.

No ano que o Beauty.AI foi lançado, cerca de 6.000 pessoas de mais de 100 países enviaram fotos na esperança de que a inteligência artificial determinasse que seus rostos eram os que mais se aproximariam do ideal de beleza humana. Entretanto, quando os resultados foram revelados, tanto criadores quanto o público ficaram incomodados ao ver que havia um fator gritante ligando os vencedores: os robôs fortemente preteriram os participantes negros. Dos 44 vencedores, quase todos eram brancos, alguns eram asiáticos e apenas um tinha pele escura. Vale ressaltar que a escolha não teve nada a ver com a distribuição de participantes por regiões. Ainda que a maioria dos participantes fosse branca, muitas pessoas negras enviaram fotos, incluindo-se grupos da Índia e da África. A controvérsia que se seguiu gerou debates sobre as maneiras pelas quais os sistemas inteligentes podem perpetuar vieses, produzindo resultados não intencionais, mas muitas vezes enviesados e até mesmo racistas [3].

O grupo de desenvolvedores e de promotores do concurso procurou prontamente comprovar que o sistema inteligente que suportava o Beauty. AI não tinha sido construído para tratar a pele clara como um sinal de beleza. Então, o que teria levado os juízes robôs a chegar à escolha de mulheres brancas como vencedoras do concurso? Para discutirmos a causa do viés no resultado do concurso é preciso lembrar que foi usado aprendizagem de máquina para gerar o modelo do que era o belo. A base de treinamento usada foi com imagens de atores e atrizes de Hollywood que em sua grande maioria, na época, eram pessoas brancas. Embora tenha se especulado sobre uma série de razões pelas quais o algoritmo favorecia os brancos, o principal problema era que os dados que o projeto usou para estabelecer padrões de atratividade não incluíam minorias em quantidades suficientes. Não houve intenção dos desenvolvedores de privilegiar os brancos, só houve um certo descuido de verificar a existência desses possíveis vieses na base de treinamento. A maneira de consertar isso é alterar a base de treinamento.

Ninguém nega a importância de um concurso de beleza e os danos que preconceitos nessa área podem causar, mas há várias outras tarefas em que esses vieses podem causar

1 http://beauty.ai/

estragos de alto impacto. Em tribunais nos EUA, já existem juízes usando sistemas inteligentes, como por exemplo o COMPAS[2], para auxiliar na tomada de decisão sobre liberdade condicional. O sistema inteligente aprende a sugerir a partir da base de casos de reincidência ao crime. Porém isso só reforça o preconceito estrutural que acaba prendendo mais negros que brancos, portanto, a base terá mais dados relacionando negros a crimes do que brancos. Preocupadas, associações de direitos humanos vêm denunciando os vieses raciais que estão aparecendo. As conseqüências neste caso são, portanto, bem mais graves do que ser preterido/preterida num concurso de beleza. Vale ressaltar que o nosso Supremo Tribunal Federal também usa IA.

Um outro exemplo de aprendizado de máquina que apresentou problemas por vieses na aquisição de dados foi o caso da Tay desenvolvido pela Microsoft em 2016. Tay foi um experimento que envolveu a utilização de aprendizado de máquina, processamento de linguagem natural e mídias sociais. Tay foi desenvolvida usando aprendizagem adaptativa, uma técnica de ponta na época e ainda muito usada. Tay foi projetado para ser um agente de conversação (*chatbot*) que aprenderia com a interação humana a dialogar naturalmente, imitando o padrão de conversação humana. Tay pedia às pessoas que enviassem selfies e ela retornava comentários divertidos, mas honestos. O objetivo era ser considerada uma adolescente que vai aprendendo e se adaptando à conversa. Aliás, um agente artificial inteligente ser confundido com um humano tem sido objetivo de todo programa de IA: passar no teste de Turing [4]

Os engenheiros da Microsoft treinaram o algoritmo de Tay em um conjunto de dados de dados públicos anônimos, juntamente com algum material pré-escrito fornecido por comediantes profissionais para dar a ele uma compreensão básica da linguagem e do que seria um comentário divertido. O plano era liberar Tay *online* e, em seguida, deixar o bot descobrir padrões de linguagem por meio de suas interações, que ela iria emular em conversas subsequentes. Conseqüentemente, seus programadores esperavam, Tay soaria exatamente como o que aprendera na Internet. A Microsoft então lançou o Tay ao público no Twitter. No início, Tay se envolveu de forma inofensiva com seu crescente número de seguidores com brincadeiras e piadas infantis. Entretanto, depois de apenas algumas horas, Tay começou a twittar coisas altamente ofensivas e agressivas, tais como: "Eu f @ #% & * # odeio feministas e todas deveriam morrer e queimar no inferno" ou "Bush fez o 11 de setembro e Hitler teria feito um trabalho melhor ..." Dentro de 16 horas de sua libertação, Tay tweetou mais de 95.000 vezes, e uma porcentagem preocupante de suas mensagens era abusiva e ofensiva. Os usuários do Twitter começaram a registrar sua indignação e a Microsoft

2 https://www.equivant.com/northpointe-suite-case-manager/

não teve outra escolha a não ser suspender a conta da Tay que virou um caso de estudo interessante de como a IA pode acabar mal rapidamente [5].

### Dados têm validade

Além da inexistência de neutralidade nos dados, uma segunda característica que deve ser levada em conta diz respeito à validade do conhecimento. Conseqüentemente, as informações contidas nas bases de dados que guiam o aprendizado da máquina podem estar datadas. Logo, a tomada de decisão num momento ou determinado contexto histórico pode ser totalmente diferente de outro, pode ser até mesmo inaceitável. Em 2018, a Amazon resolveu ampliar seu processo de recrutamento. Como sabia que iria receber milhares de currículos, decidiu investir em um sistema inteligente que faria uma pré-seleção dos currículos. Para treinar o sistema, ela contou com a vasta base dados dos seus funcionários. O desejo dos projetistas e responsáveis pelo recrutamento era contratar pessoas que se ajustassem bem ao estilo da empresa. Entretanto, o resultado do processo seletivo foi parar nas páginas dos jornais. Nenhuma mulher foi pré-selecionada. E mais, nenhum homem que tivesse estudado em universidade com nome de mulher foi selecionado. A empresa pediu desculpas e disse que não era a sua intenção. O que não foi levado em conta pelos projetistas e desenvolvedores é que a presença de mulheres na área de computação e, mesmo no comércio eletrônico, é recente. A base de dados de funcionários da empresa era majoritariamente masculina. Portanto, os funcionários bem-sucedidos ao longo da história da empresa eram em sua grande maioria homens e foi isso que o sistema aprendeu. A Amazon não está sozinha no uso de sistemas que, mesmo sem pretenderem, acabam realizando recrutamento preconceituoso, como é o caso do HireVue[3] e PredictiveHire[4] .

Para Hsu [6], empresas como a Amazon tem feito esforço para afirmar que, com design e com treinamento cuidadosos de seus modelos de IA, eles serão capazes de abordar especificamente várias fontes de viés sistêmico em uma prospecção de recrutamento. O autor insiste que esta não é uma tarefa simples, já que algoritmos de IA têm sido questionados e acusados de injustiça em relação a gênero,à raça e à etnia. As estratégias adotadas por essas

3 https://www.hirevue.com/
4 https://www.predictivehire.com/

empresas incluíram identificação de limpeza de informações de aplicativos, contando com entrevistas anônimas e testes de conjunto de habilidades, além de oferecer o serviço de ajuste do texto do trabalho postagens para atrair um campo de candidatos possível.

### Dados podem carregar vieses escondidos

A terceira característica a ser considerada é que os vieses e os preconceitos podem estar escondidos nos dados, o que torna a tarefa de identificá-los mais difícil. Obermeyeret e colegas [7] publicaram uma pesquisa mostrando o viés racial em sistemas inteligentes na área da saúde. Segundo o estudo, uma grande empresa de seguro saúde norte-americana tinha o objetivo de reduzir seus custos. Numa primeira análise, notou que seus assegurados com doenças crônicas graves davam entrada com muita freqüência nos serviços de emergência e utilizavam muito os centros de tratamento intensivo (CTI). Como se tratam de procedimentos dispendiosos, a empresa resolveu então oferecer tratamentos preventivos. Numa avaliação interna, a ação seria vantajosa para todos, já que a empresa reduziria os custos e os pacientes teriam mais chances de sobrevida. Porém, tais tratamentos também refletiam em custos. Para oferecer esses tratamentos preventivos com mais eficiência, a empresa resolveu fazer uma triagem e oferecer apenas para os casos considerados mais críticos. A definição do que era caso crítico estava associada a quanto o paciente usava do sistema de saúde. Quanto mais ele usava o sistema de saúde em consultas e internações, quanto mais ele parava no CTI ou na emergência, mais ele deveria precisar de cuidados.

Com isso, pacientes com registros clínicos semelhantes eram considerados mais críticos, se usassem mais os serviços de saúde oferecidos pelo plano. O sistema aprendeu esse padrão da enorme base de dados dos clientes do seguro. Para evitar qualquer viés racial, foram retiradas da base quaisquer informações nesse sentido. O sistema ficou em uso de 2013 a 2015. Ao investigar o resultado, constatou-se que as pessoas que foram escolhidas para receberem o tratamento preventivo eram em sua grande maioria brancas. Mas se a informação sobre raça tinha sido retirada da base, por que o sistema estava privilegiando os brancos a receberem o tratamento preventivo? Sem nenhuma teoria que pudesse corroborar a maior fragilidade dos brancos, uma investigação aprofundada concluiu que não se tratava disso. Os negros usavam menos o seguro saúde porque eram mais pobres. Eles não podiam pagar a contrapartida (franquia ou *deductible*) exigida pelo seguro. A cada visita médica, o paciente sempre paga algum percentual da conta. Além disso, o que a pesquisa mostrou foi uma certa instabilidade empregatícia dos seus assegurados negros que não queriam perder dias de trabalho indo ao médico com um medo justificável na época de perder o emprego. Portanto a base de dados não tinha muitos dados das visitas médicas de assegurados negros que permitissem ao sistema

identificar o padrão de gravidade que os incluíssem na lista de pacientes a receberem o tratamento preventivo.

Esse é o resultado de desigualdades arraigadas que significam que os negros vão a menos consultas médicas e, quando o fazem, os médicos prescrevem, em média, menos medicamentos e solicitam menos exames. Portanto, ao olharmos uma base temos que ter o cuidado de investigar os possíveis vieses escondidos. Exemplos como esses têm fomentado pesquisas ao redor de todo o mundo.

## Considerações Finais

O que os casos discutidos neste artigo nos revelam é que é preciso reconhecer e discutir as distorções que o emprego de técnicas de inteligência artificial não só exacerba, mas perpetua, como vieses raciais e desigualdades. Dados não são neutros. Eles registram decisões humanas que são processos de escolhas e tais escolhas podem estar impregnadas de preconceitos. Um sistema inteligente eficiente aprende dos dados tais preconceitos e os consolida. Mais grave ainda é que as decisões vindas da máquina vêm revestidas de mérito pela performance nas métricas matemáticas de acurácia e precisão, o que lhes confere uma pretensa aura de imparcialidade. É importante que os desenvolvedores entendam sua responsabilidade no desenvolvimento de sistemas inteligentes que sejam éticos para não reproduzirem em larga escala, através de algoritmos e redes de Inteligência Artificial, os vieses que os dados carregam. Entender o contexto na geração dos dados e no uso atual, estressar o sistema para identificar grupos que possam ser prejudicados com as respostas e criar sistemas que sejam capazes de explicar suas respostas são algumas das atitudes que privilegiariam a ética nos sistemas. Além disso, o cidadão tem que estar atento ao seu direito de resguardar sua privacidade e mesmo a propriedade de seus dados e deve ser incentivado a pelo menos a ter a consciência da utilização de uma enorme massa de dados e de transações financeiras e do seu dia a dia para fins nobres de segurança e de saúde. Cabe ao cidadão, portanto, verificar se o que foi aprendido está de acordo com os padrões éticos da sociedade. É preciso exigir das empresas e dos governos que seus sistemas inteligentes sejam transparentes e auditáveis.

Há ainda problemas éticos que surgem pelo uso indiscriminado da IA. O sistema DeepGestalt [8] foi desenvolvido para identificar problemas genéticos de saúde através da análise de características faciais. Existe um benefício fantástico para tratamentos preventivos de desenvolvimento de doenças. Por outro lado, tal sistema pode ser usado para discriminar candidatos no recrutamento e contratação de funcionários ou para precificarem o seguro saúde. Há ainda sistemas que usam a IA, mas já começam com objetivos eticamente errados como é o caso do sistema Gaydar [9] que identifica se uma pessoa é homossexual a partir da análise de fotos. Esse é um uso intrinsicamente aético.

Embora um progresso significativo

tenha sido feito em últimos anos na área técnica e multidisciplinar de pesquisa, mais investimento nesses esforços serão necessários. Os governos, centros de pesquisa e mesmo líderes empresariais também podem ajudar a apoiar o progresso da IA ética tornando os dados que alimentam os sistemas disponíveis ao escrutínio externo de pesquisadores.

Para um uso consciente e com menos vieses, faz-se indispensável uma abordagem multidisciplinar, que inclua especialistas em Ética, cientistas sociais, e especialistas que melhor entendem as nuances de cada área de aplicação de Inteligência Artificial.


Referências

1. J. Frith and M. Saker. It Is All About Location: Smartphones and Tracking the Spread of COVID-19. Soc. Media Soc., vol. 6, no. 3, pp. 2–5, 2020, doi: 10.1177/2056305120948257.
2. Q. V. Pham, D. C. Nguyen, T. Huynh-The, W. J. Hwang, and P. N. Pathirana. Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts. IEEE Access, vol. 8, no. April, pp. 130820–130839, 2020, doi: 10.1109/ACCESS.2020.3009328.
3. S. D'Souza and B. Mehta. Defining a Sandbox for Responsible AI. SSRN Electron. J., pp. 2–6, 2018, doi: 10.2139/ssrn.3255075.
4. Saygin, Ayse Pinar, Ilyas Cicekli, and Varol Akman. Turing test: 50 years later. Minds and machines 10.4 (2000): 463-518
5. G. Neff, and P. Nagy. Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay" International Journal of Communication 10 (2016): 17.
6. J. Hsu. Can AI hiring systems be made antiracist? Makers and users of AI-assisted recruiting software reexamine the tools' development and how they're used-[News]. IEEE Spectrum 57.9 (2020): 9-11.
7. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. Science (80-. )., vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
8. Y. Gurovich, et al. DeepGestalt-Identifying rare genetic syndromes using deep learning. arXiv preprint arXiv:1801.07637 (2018).
9. C. Jernigan and BFT Mistree. Gaydar: Facebook friendships expose sexual orientation. First Monday (2009).



**ANA CRISTINA BICHARRA GARCIA** é Professora Titular do Departamento de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO), com doutorado pela Stanford University, Califórnia, EUA (1992). É pesquisadora do CNPq, vice-coordenadora da comissão especial de sistemas colaborativos (CESC-SBC) e coordenadora do programa de pós-graduação em informática (PPGI) da UNIRIO. Atua nas áreas de Inteligência Artificial e Inteligência Coletiva. Seus interesses de pesquisa atuais focam em modelos de explicabilidade em sistemas inteligentes, ética em IA, privacidade de dados e legado digital.