Capítulo 35

Questões éticas em IA e PLN

Maria das Graças Volpe Nunes
Tayane Arantes Soares
Mariza Ferro

Publicado em: 26/09/2023
Atualizado em: 20/11/2024

Toda civilização, coletividade ou sociedade surge a partir do compartilhamento de necessidades, bens e valores comuns. A sobrevivência e a prosperidade de uma sociedade dependem de algum tipo de mediação das diferenças e da regulação do comportamento de seus integrantes. A mediação das diferenças e a regulação da conduta de indivíduos partem de pressupostos. A ética refere-se ao comportamento de indivíduos na tomada de decisões e na sua responsabilização por elas, frente aos valores compartilhados pela sociedade em que vivem.

A partir do momento em que os sistemas de inteligência artificial (IA) passam a fazer parte da sociedade, interagindo com humanos e mimetizando seus comportamentos, tomando decisões com certo grau de autonomia e eventualmente colocando pessoas ou sociedades em risco, problemas de natureza ética naturalmente emergem. Nesse contexto, tem havido uma preocupação crescente com as implicações éticas dos atuais sistemas inteligentes, e a sociedade acadêmica de IA tem se movimentado para alertar e promover mudanças para minimizá-los ("AI and Ethics", 2023; Coeckelbergh, 2020).

35.1 Ética em IA

Atualmente (2024), a principal tecnologia de IA para dotar seus programas com inteligência caracteriza-se por fornecer, a algoritmos criados para aprender, grandes quantidades de dados sobre aquilo que deve ser aprendido, ou seja, sobre um conceito ou uma tarefa. E, conforme já discutido no Capítulo 13, se esses dados não forem coletados de maneira criteriosa, podem conter vieses que acabam por provocar comportamentos indesejáveis, incorretos ou inadequados. Isso ocorre muitas vezes por terem sido treinados com dados desbalanceados e sem curadoria, ou por terem aprendido correlações entre os dados que ou são irrelevantes para o conceito que se quer ensinar, ou carregam algum viés indesejado.

Um exemplo concreto de vieses algorítmicos é relatado no documentário "*Coded Bias*" (Kantayya, 2020). A cientista da computação Joy Buolamwini, uma mulher negra, durante a sua pesquisa sobre softwares de visão computacional no MIT (*Massachusetts Institute of Technology*), não conseguia ter seu rosto reconhecido pelo software no qual estava trabalhando. Somente após colocar uma máscara facial branca que o sistema reconheceu a máscara como sendo um rosto. As pesquisadoras Joy Buolamwini e Timnit Gebru

constataram que 79,6% dos dados de treinamento desse software eram compostos por pessoas de pele clara (Buolamwini; Gebru, 2018).

Esse fato evidencia o problema estrutural que permeia a criação de ferramentas de IA, tendo em vista a pouca ou nenhuma reflexão por parte de empresas e de pessoas que desenvolvem essas soluções sobre os impactos sociais que essas ferramentas podem causar, além de pouco envolvimento da sociedade no desenvolvimento dessas soluções tecnológicas (Hora, 2021). O'Neil (2021) relata vários outros exemplos de consequências negativas de se utilizar algoritmos de aprendizado de máquina (AM) em tomada de decisões.

Casos de discriminação de raça são frequentes. A ferramenta *Google Fotos* foi acusada de rotular a imagem de um casal negro como "gorilas", e a do *Flickr* rotulou fotos de pessoas negras como "macaco" (Cruz, 2021). A pesquisa de Buolamwini; Gebru (2018) também revelou vieses de raça e de gênero em serviços de IA de empresas como Microsoft, IBM e Amazon. O então Twitter, em 2020, foi denunciado por priorizar rostos de pessoas brancas na exibição de imagens publicadas pelos usuários (INFOBASE, 2021).

A IA também já foi acusada de impulsionar o ódio às minorias e influenciar os resultados de eleições (Cavaliere; Romeo, 2022), explorar fraquezas psicológicas e orientar decisões (Sartori; Theodorou, 2022), causando problemas como a intensa polarização social e ameaças aos princípios democráticos e aos direitos humanos (Artificial intelligence and human rights., 2021; Empoli, 2019).

Outra característica importante desses algoritmos que aprendem (ao menos na abordagem mais utilizada no ano de 2024, que são as redes neurais artificiais) é que aquilo que aprendem não serem recuperáveis de forma compreensível a nós, ou seja, é impossível recuperar exatamente qual conhecimento foi apreendido pela máquina. Aferimos seu conhecimento apenas pelo seu comportamento numa determinada tarefa. Nesse sentido, dizemos que são obscuros, verdadeiras caixas-pretas, ou seja, não explicáveis. Essa característica indesejável deriva da tecnologia chamado *Deep Learning* (Goodfellow et al., 2016), que nada mais é do que um modelo especial de redes neurais. Desempenham muito bem, é verdade, mas não conseguimos entender como o fazem. Desse modo, a impossibilidade de se justificar, aliada à presença cada vez mais sentida desses sistemas no nosso cotidiano, tem gerado um sentimento de insegurança e desconfiança.

O ChatGPT, da OpenAI (OpenAI, 2022), de enorme repercussão no final de 2022, rapidamente teve sua reputação abalada devido à incapacidade de referenciar com exatidão a fonte de suas respostas (até porque, como foi mencionado anteriormente, é extremamente complicado recuperar com precisão o conhecimento apreendido pelo modelo a partir dos dados de treinamento (Heikkilä, 2021)). Consequentemente, torna-se desafiador determinar a fonte exata das respostas geradas, uma vez que o modelo atua essencialmente como um gerador de palavras prováveis com base em uma entrada inicial. Para algumas situações em que essas respostas determinariam decisões ou teriam consequências importantes, a falta de confiança no sistema pode ter afastado alguns usuários. A despeito disso, no entanto, este e outros *chatbots* do mesmo tipo têm sido rapidamente adotados por usuários em suas tarefas de trabalho, seja como uma ferramenta poderosa de criação e edição de textos, seja como uma perigosa fonte de informações muitas vezes incorretas. Para saber mais sobre a tecnologia do ChatGPT, sugerimos a leitura de Capítulo 20.

A IA remete a problemas éticos na medida em que são construídos artefatos (sistemas, robôs) que interagem com humanos (de forma direta ou indireta, visível ou ubíqua) e que o resultado dessas interações podem afetar, com algum risco, esses usuários ou a sociedade como um todo. Considerando-se que sistemas inteligentes tendem a ultrapassar barreiras físicas, sociais, temporais e culturais (tendo em vista que são usados em vários lugares

do mundo), devemos nos lembrar que seus desenvolvedores sempre estarão inseridos num contexto cultural e moral específicos, o que pode entrar em conflito com os valores éticos de sociedades usuárias distintas.

A fim de evitar problemas dessa natureza, em uma sociedade cada vez mais interativa com máquinas de IA, é fundamental investigar maneiras de construir esses artefatos de maneira responsável (Russel, 2019). Caso contrário, essas novas tecnologias continuarão a perpetuar pontos de vista hegemônicos, reforçando e codificando preconceitos e vieses humanos que ainda lutamos para combater (Bender et al., 2021). Nina da Hora, cientista da computação brasileira e pesquisadora na área de Pensamento Computacional, ressalta que, muitas vezes, a busca global pela ética em IA, por ser baseada na tentativa de manter as tecnologias que estão causando problemas, não aprofunda o entendimento e a investigação dos problemas enfrentados pelas pessoas afetadas por essas tecnologias. Segundo a pesquisadora, é necessário ir além dos aspectos técnicos ao buscar um desenvolvimento ético de sistemas de IA, e investigar também o impacto dessas novas tecnologias na vida das pessoas envolvidas (Hora, 2022).

Existem diversas recomendações e tentativas de regulação dos sistemas de IA ao redor do mundo, sendo que os modelos regulatórios de EUA, China e União Europeia destacam-se como as três principais propostas, apesar de haver outros modelos competitivos, como o do Canadá, por exemplo. Nos EUA o foco tem sido evitar excessos regulatórios, caminhando para cultura da autorregulação, em que princípios são declarados e segui-los é responsabilidade das empresas. Porém, para isso há medidas concretas e rígidas sendo estabelecidas e se caminha para impor restrições e exigência de deveres para as grandes empresas. A China mantém um modelo de regulação mais impositivo com maior controle estatal, concentrado em um subconjunto de aplicações de IA, mas que trabalha para atualizar a sua regulação para uma nova lei mais abrangente. Entre os dois extremos, a União Europeia, pioneira na discussão sobre IA (Commission, 2021) aprovou em maio de 2024 a primeira Lei da Inteligência Artificial codificada (*Artificial Intelligence Act*), uma lei baseada no risco, ou seja, quanto maior o risco de danos à sociedade, mais rigorosa será a lei. Embora prevista para entrar em vigor apenas em 2026, algumas proibições se anteciparão: fica proibido usar sistemas para pontuação social, policiamento preditivo e coleta indiscriminada de imagens faciais disponíveis na internet ou capturadas por câmeras de circuito interno. Por exemplo, o uso de vigilância biométrica pelo governo se restringirá a casos que envolvam crimes, terrorismo e busca de suspeitos de crimes graves[1].

No Brasil, o mais recente projeto de lei (PL 2338/2023) encontra-se, em 2024, em tramitação na Câmara dos Deputados e no Senado Federal. O projeto estabelece "normas gerais de caráter nacional para o desenvolvimento, implementação e uso responsável de sistemas de inteligência artificial (IA) no Brasil, com o objetivo de proteger os direitos fundamentais e garantir a implementação de sistemas seguros e confiáveis, em benefício da pessoa humana, do regime democrático e do desenvolvimento científico e tecnológico"[2]. Pode-se dizer que todas as questões aqui levantadas são, na letra da lei, contempladas por esse projeto. No entanto, considerando-se o domínio absoluto de poucas e grandes *Big Techs* americanas e a complexidade e subjetividade envolvidas, o grande desafio será fiscalizar o cumprimento da lei.

Até 2023 existiam pelo menos 200 documentos com princípios destinados a fornecer orientações normativas em relação aos sistemas baseados em IA em vários países (Corrêa et al., 2023), nos quais destacam-se os princípios promovidos pela OCDE (Organização para

[1] https://artificialintelligenceact.eu/ai-act-explorer/

[2] https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1720062232582&disposition=inline

a Cooperação e Desenvolvimento Econômico) para classificação e avaliação de sistemas de IA, que fomenta a universalização de critérios para políticas de IA (OECD, 2022). Vale ainda destacar o documento da UNESCO, aprovado em novembro de 2021, reconhecendo os impactos positivos e negativos da IA nas sociedades e recomendando que os Estados-membros tomem providência quanto à violação de direitos (UNESCO, 2022). O objetivo é sempre recomendar princípios para que os sistemas de IA sejam confiáveis, desenvolvidos e utilizados para o bem da humanidade e do planeta e para preservar os valores por meio da proteção, promoção e respeito aos direitos humanos fundamentais, à liberdade e à igualdade.

A utilização de sistemas de IA pode afetar negativamente vários direitos fundamentais estabelecidos pela Declaração Universal dos Direitos Humanos, adotada e proclamada pela Assembleia Geral das Nações Unidas (UNICEF, 1948) ou por instrumentos particulares de cada país, como a Constituição Brasileira ou a Carta dos Direitos Fundamentais da União Europeia. Os direitos fundamentais são inerentes a todos os seres humanos, independentemente da sua raça, sexo, nacionalidade, etnia, idioma, religião ou qualquer outra condição. Os direitos humanos incluem o direito à vida e à liberdade, liberdade de opinião e expressão, o direito ao trabalho e à educação, entre outros.

Entre os princípios mais comuns que norteiam as regulações e as recomendações para o desenvolvimento e o uso da IA ética e confiável estão a (1) justiça, diversidade e não discriminação, (2) transparência e explicabilidade, (3) robustez técnica e segurança, (4) privacidade e proteção de dados, (5) responsabilidade e prestação de contas[^cap-etica-3].

1. Os princípios da justiça, diversidade e não discriminação estão intimamente ligados à promoção da justiça social e a salvaguardar a equidade e a não discriminação de qualquer tipo (gênero, raça, cor, nacionalidade, religião, língua, idade, opinião política etc.), em conformidade com o direito internacional. O objetivo é garantir a distribuição igual e justa de benefícios e custos e garantir que indivíduos e grupos estejam livres de preconceitos, injustiças, discriminação e estigmatização. Ainda, minimizar e evitar reforçar ou perpetuar resultados discriminatórios ou tendenciosos (enviesados), ao longo do ciclo de vida dos sistemas de IA (Smith; Rustagi, 2020). Se preconceitos e injustiças não puderem ser evitados, os sistemas de IA podem aumentar a desigualdade social. Além disso, o uso de sistemas de IA nunca deve induzir as pessoas a serem enganadas ou prejudicar sua liberdade de escolha.

   Dependendo da forma como é criada e utilizada, a IA tem potencial para criar e/ou reforçar vieses humanos. O viés pode entrar no desenvolvimento e uso de um sistema de IA, especialmente por meio do uso dos algoritmos de aprendizado de máquina durante a geração, a coleta, a rotulagem e o gerenciamento dos dados com os quais o algoritmo aprende; mas também pode ser introduzido durante o design e a avaliação dos algoritmos (Smith; Rustagi, 2020). Já existem muitos exemplos do uso de sistemas que utilizam AM, os quais, com base nos dados que recebem, têm apresentado resultados tendenciosos, imprecisos e injustos, os quais representam riscos imensos para indivíduos e empresas.

   São diversas as situações de discriminação de raça e de vieses e discriminação aos grupos minoritários ou culturas, principalmente com o uso de algoritmos de reconhecimento facial ou manipulação de imagens. Por exemplo, a dificuldade em reconhecer rostos de pessoas negras, como os exemplos mencionados no início deste capítulo, com a classificação automática de fotos de pessoas negras como “gorilas”; aplicativo que “desnudava” mulheres mostrando como as *deepfakes* prejudicam os

mais vulneráveis (REVIEW, 2022); aplicativos que transformam fotos em caricaturas onde os avatares das mulheres, especialmente orientais, são "pornificadas", enquanto o dos homens são astronautas, exploradores e inventores (Heikkilä, 2022). Além dos casos das mídias sociais, o Brasil vem sofrendo uma profusão de denúncias com o uso de sistemas de reconhecimento facial que levaram a abordagens policiais e até prisões. A Rede de Observatórios da Segurança monitorou, entre março e outubro de 2019, os casos de prisões e abordagens com o uso de reconhecimento facial em cinco estados brasileiros e revelou que 90,5% dos presos por monitoramento facial no Brasil são negros (Nunes, 2019).

Há duas formas para minimizar esses vieses: fazer com que os dados de treinamento reflitam fielmente o universo de situações a que o sistema final será exposto, e detectar, ainda em período de testes, os desvios prováveis e eliminá-los antes de colocar o sistema em uso. Considerando a natureza da tecnologia mais comum hoje em IA – o treinamento de redes neurais – as duas formas são de difícil execução. Quer seja porque nem sempre todos os dados são acessíveis, quer seja porque os testes são incapazes de prever todas as possibilidades, dada a complexidade da tarefa a ser realizada pelo sistema.

2. Os princípios da transparência e explicabilidade são fundamentais para desenvolver e manter a confiança dos usuários nos sistemas de IA. Isso significa que os processos precisam ser transparentes, ou seja, o objetivo dos sistemas de IA deve ser comunicado abertamente e o usuário deve saber que está em contato com um produto ou serviço fornecido diretamente ou com o auxílio de sistemas de IA. Além disso, as decisões tomadas pelo sistema – na medida do possível – devem ser explicáveis aos afetados direta ou indiretamente. Nem sempre é possível explicar por que um modelo gerou uma determinada saída ou decisão (e qual combinação de fatores de entrada contribuiu para isso). Esses casos são chamados de algoritmos "caixa-preta" como os já mencionados algoritmos de redes neurais e suas redes neurais profundas. A comunidade de IA já se mobiliza em direção a tornar seus sistemas mais compreensíveis para seus usuários. A IA explicável (*Explainable AI*, *XAI*) é uma recente área de pesquisa que tem como objetivo propor processos e métodos para tornar os recentes sistemas de aprendizado de máquina mais compreensíveis e confiáveis. Não é tarefa fácil, mas, juntamente com esforços para combinar aprendizado de máquina com métodos simbólicos de representação de conhecimento, é esperado que testemunharemos avanços nessa área.

3. Um componente crucial para alcançar uma IA confiável é a robustez técnica, que está intimamente ligada ao princípio da prevenção de danos. A robustez exige que os sistemas de IA se comportem conforme o planejado, sejam desenvolvidos com uma abordagem preventiva aos riscos, minimizando danos não intencionais e inesperados e evitando danos inaceitáveis. Além disso, vulnerabilidades a ataques (riscos de segurança) devem ser evitadas e eliminadas durante o ciclo de vida dos sistemas de IA para garantir proteção e segurança humana e ambiental.

4. A privacidade e a proteção de dados devem ser respeitadas, protegidas e promovidas ao longo do ciclo de vida dos sistemas de IA. É importante que os dados destinados aos sistemas de IA sejam coletados, utilizados, compartilhados, arquivados e apagados de modo compatível com os marcos jurídicos nacionais, regionais e internacionais relevantes. Por exemplo, na legislação brasileira, a LGPD (Lei Geral de Proteção aos Dados) estabelece diretrizes para o uso dos dados pessoais (LGPD, 2018). A lei

similar europeia, GDPR, vai além e trata também do direito de que dados pessoais sejam apagados das bases de dados a qualquer momento, tratado como o direito ao esquecimento pelos modelos de IA.

O uso de dados pessoais ocorre em diferentes contextos. Exemplos incluem o uso de dados para a identificação de supostas emoções para análise do comportamento do usuário. No caso de um cliente, por exemplo, para criar publicidade direcionada durante as compras com foco nos produtos ou em sua disposição na loja virtual, sem a transparência desejada. A ViaQuatro, empresa que tem a concessão da linha 4-amarela do metrô de São Paulo, foi processada pelo Instituto Brasileiro de Defesa do Consumidor por usar câmeras que coletavam dados referentes às "emoções" dos passageiros, e que seriam usados pela companhia, sem o consentimento dos passageiros. "O sistema inteligente conseguiria identificar se o passageiro está feliz, insatisfeito, surpreso e neutro. Além disso, detectaria o gênero e a faixa etária das pessoas. Os dados capturados seriam usados para a empresa fazer a gestão de seu conteúdo institucional e até de anúncios publicitários" (Cruz, 2018). Mais um exemplo de uso sem transparência e desrespeito à privacidade e a proteção dos dados pessoais.

5. A responsabilização e a prestação de contas complementam os princípios acima e estão intimamente ligadas ao princípio da justiça ao tentar garantir mecanismos que determinem as responsabilidades éticas e jurídicas pelas decisões e ações de alguma forma baseadas em um sistema de IA e seus resultados, antes e depois do seu desenvolvimento, implantação e uso.

   Enquanto se aguarda uma regulamentação direcionada aos sistemas de IA no Brasil, a LGPD tem sido usada para responsabilizar alguns abusos, como no caso citado, da ViaQuatro. A lei diz que a captação de dados sensíveis, como os biométricos, precisa de consentimento do usuário, e a finalidade da coleta também deve ter propósitos legítimos e comunicados aos titulares dos dados. Além disso, o direito à informação dos consumidores está consagrado como um princípio fundamental ao abrigo do Código do Consumidor.

   Porém, em casos não cobertos por outras leis, a responsabilização por danos causados por sistemas de IA ainda carece de lei própria. O problema é que o avanço tecnológico ocorre em ritmo muito mais rápido do que aquele da política e da justiça. Enquanto se discutem as leis para regular esta IA de hoje, ela continua avançando e modificando nossa forma de interagir com ela e por meio dela, e antes que seja regulamentada por leis, ela pode ter se tornado outra.

A grande questão é se, um dia, um sistema de inteligência artificial estará programado para avaliar adequadamente as informações recebidas e as possíveis consequências que suas ações são capazes de causar ao ambiente e aos seres à sua volta. Será possível programá-lo para embasar suas decisões à luz de valores humanos a fim de exibir um comportamento ético?

Essa possibilidade esbarra em várias dificuldades, como a definição dos valores responsáveis por um comportamento ético, sua representação (seja de forma explícita ou por meio de exemplos), seu processamento por um algoritmo e sua incorporação por um sistema de inteligência artificial.

35.2 Ética em PLN

Não é coincidência que as discussões sobre a regulação de IA se intensificaram após o lançamento do ChatGPT. Os sistemas de PLN, ou seja, aqueles em que a língua tem protagonismo, estão entre aqueles que mais expõem suas semelhanças conosco, suas fragilidades na compreensão da língua humana, seus perigos quando interagem conosco em cenários sensíveis - como para fazer diagnósticos ou pretensas terapias, por exemplo -, os vieses culturais e morais de seus dados, entre outras coisas. O PLN assume, assim, um papel importante no âmbito das questões éticas relacionadas aos sistemas de IA. A fim de evitar vários problemas, é necessário tratar a língua de forma ampla, considerar suas variações, suas mudanças, seu papel na comunicação humana e na sociedade; tudo isso pode nos auxiliar a construir tecnologias melhores e mais inclusivas (Bender, 2020).

Muitas línguas até então desfavorecidas de recursos tecnológicos, como o português, e também línguas minoritárias, têm se beneficiado com sua inserção no mundo digital. No entanto, ainda temos um longo caminho a percorrer. Segundo Emily Bender, 90% das línguas do mundo e suas variedades usadas por mais de um bilhão de pessoas têm pouco ou nenhum suporte em termos de tecnologia linguística, reforçando a ideia de que essas novas tecnologias apresentam um potencial excludente (Bender, 2020).

Nesse sentido, a comunidade de PLN deve ter em mente que vivemos em um mundo linguisticamente diverso e que não é razoável aceitar o inglês, idioma dos dados de treinamento da maioria dos modelos de língua, como representativo de toda variedade linguística e cultural existente. Ao trabalharmos com PLN, não podemos esquecer que linguagem e poder estão atrelados, que a linguagem cria o nosso mundo e molda nossa realidade (HALLIDAY; MATTHIESSEN, 1999). Portanto, temos o desafio de romper com esse monopólio linguístico e desenvolver modelos a partir de dados coletados de forma responsável, validados, balanceados e livres de vieses.

É fato que o PLN tem se beneficiado muito com o uso de aprendizado de máquina. Muitas barreiras foram transpostas ao representar alguns fenômenos linguísticos por meio de exemplos. Para as tarefas clássicas de PLN – *taggers*, *parsers*, reconhecedores de entidades nomeadas, entre outras – os sistemas construídos por AM parecem cada vez melhores à luz de avaliações padronizadas. O problema que se coloca é o uso futuro desses sistemas, suas combinações e suas aplicações fora de qualquer controle (Chandran, 2023).

Nos sistemas de IA mais recentes, a competência linguística é adquirida por meio de treinamento com *corpora* muito grandes, gerando um modelo de língua, ou seja, um sistema capaz de prever qual(is) palavra(s) deve(m) seguir a última palavra vista. São os chamados LLM (*Large Language Models*), apresentados em Capítulo 17. O ChatGPT é um exemplo muito conhecido dessa tecnologia.

Se considerarmos que os *corpora* de treinamento de grandes modelos de língua tendem a ser compostos por uma quantidade massiva de dados linguísticos coletados na internet, e que o acesso à internet é desigual, os dados de treinamento têm grandes chances de não serem representativos e não levarem em conta a diversidade cultural e linguística existentes - desconsiderando aqui o uso não autorizado de dados pessoais. Conforme discutido no Capítulo 34, manifestações carregadas de ofensas, preconceitos, discriminação, posturas antiéticas em geral são eventualmente reproduzidas nos textos gerados por esses modelos (Perrigo, 2023), e acabam se perpetuando no modelo gerado.

Além disso, a popularização de modelos de língua, como o ChatGPT, tem suscitado questões importantes no âmbito da ética em PLN, especialmente no que diz respeito à propriedade intelectual e aos direitos autorais. Embora esses dados estejam disponíveis

publicamente, uma preocupação fundamental está relacionada à forma como os dados são coletados para o treinamento desses modelos, considerando se as pessoas que produziram esses dados tinham ciência de que suas postagens textuais poderiam ser utilizadas como insumos para modelos de linguagem (Alisson, 2023). Ilustra esse ponto a localização labiríntica, nos produtos da Meta, do formulário que permite ao usuário coibir o uso de seus dados no treinamento de algoritmos [3]. A pergunta que emerge é: como podemos garantir que princípios éticos de transparência, proteção de dados e consentimento serão respeitados nesse processo de coleta?

35.3 Modelos de língua como fonte de conhecimento?

Modelos de língua nos surpreendem ao possibilitarem a geração de textos coerentes, muitas vezes indistinguíveis de textos produzidos por seres humanos. No entanto, esses textos não passam de sequências de palavras prováveis estatisticamente em um dado idioma que foram "cuspidas" pelo modelo a partir de alguma entrada textual. Os modelos em si não entendem os textos gerados. Os modelos apenas apreendem, a partir do *corpus* de treinamento, os padrões (combinações frequentes) linguísticos derivados dos dados, e reproduzem, como bons papagaios estocásticos, esses padrões em novas saídas (Bender, 2023).

Nesse sentido, se a língua é apreendida a partir de um *corpus*, nos modelos de línguas, as características desse *corpus* são determinantes para a qualidade linguística do que será gerado pelo sistema (Capítulo 13). Isso soa óbvio, mas, em se tratando de língua, há uma outra consequência. Em sistemas conversacionais, como os *chatbots*, a linguagem produzida por um sistema tem um efeito: ela atenderá a alguma expectativa do usuário, que pediu uma informação, ou uma sugestão, ou se queixou de algo, ou quer simplesmente dialogar. Não basta, portanto, que a expressão linguística cumpra todos os requisitos de ortografia, gramática, coesão e coerência. É preciso atender a outros critérios. Não é rara a geração de uma expressão linguística correta e elegante, com um conteúdo ou uma informação incorreta ou enviesada pelos dados. Detectar essa imprecisão, no entanto, pode não ser tão fácil. Um interlocutor do *chat*, impressionado pela boa forma do texto, pode aceitar como verdade, sem questionar, o conteúdo expresso por ela. Acontece que um modelo de língua não é capaz de preencher os requisitos relativos à autenticidade e veracidade das expressões que gera. De fato, trata-se de uma ferramenta incrível de geração de texto sendo usada para um fim para o qual não foi projetada. Como são capazes de gerar uma infinidade de expressões linguísticas, tem-se a impressão de que, de fato, têm domínio em várias áreas de conhecimento e tarefas. A consequência é que suas "alucinações" podem ser confundidas com novas "verdades", oferecendo um risco enorme à sociedade, na medida em que a crença nessas verdades pode levar a comportamentos imprevisíveis. Percebe-se aí o perigo de se utilizar um sistema impróprio como se fosse um "gerador de conhecimento".

Tão logo foi disponibilizado o ChatGPT, em 2022, as consequências desse cenário têm sido discutidas por vários setores das sociedades em todo o mundo, incluindo o Brasil. Já se prevê mudanças no trabalho em toda sorte de setores que usam informações para tomada de decisão, bem como aqueles que têm a redação de textos como atividade relevante. Incluem-se, portanto, o jornalismo, a educação formal, a pesquisa, o direito, apenas para citar alguns.

Se mal gerenciadas, essas transformações do mercado de trabalho podem ter custos econômicos e sociais significativos. Segundo a OECD (2023), 27% dos empregos podem ser

[3] https://www.facebook.com/help/contact/510058597920541

impactados por automação com tecnologias de IA, e com o advento dos modelos generativos, certamente esses impactos aumentarão, para uma ampla gama de categorias de empregos. Para melhor proteção dos trabalhadores nos próximos anos, a OECD (2023) recomenda, entre outras medidas, que os governos garantam treinamento para lidar com a IA. A UNESCO (UNESCO, 2019) também enfatiza a necessidade de desenvolvimento de valores e habilidades para a vida e para o trabalho na era da IA e recomenda medidas para melhorar a literacia em IA em todas as camadas da sociedade. Capacitar a população com habilidades relacionadas à IA pode melhorar sua empregabilidade e prepará-los para as demandas deste mercado de trabalho em constante evolução. Além disso, compreendemos ser muito importante iniciativas para ampliar a educação em IA para públicos sub-representados em um esforço para aumentar a diversidade da força de trabalho.

Temos testemunhado que sociedades cada vez mais tecnológicas suscitam muitas questões de natureza ética. A velocidade com que os sistemas computacionais evoluem – em especial, os ditos inteligentes – tem nos mostrado que precisamos nos antecipar, de alguma forma, aos riscos que eles podem representar. Para alcançarmos desenvolvimento e utilização ética e responsável de sistemas de IA, precisamos contar com esforços coletivos e transdisciplinares, além de um diálogo constante entre governo, empresas, cientistas, especialistas e sociedade em geral. Nesse sentido, é fundamental promover debates mais amplos e plurais sobre os impactos dessas novas tecnologias, a fim de pensarmos de forma conjunta aplicações positivas dessas ferramentas em nossa sociedade.

Referências

AI and Ethics. Springer, 2023. Disponível em: <https://link.springer.com/journal/43681/volumes-and-issues>. Acesso em: 7 abr. 2023

ALISSON, S. **Their god is not our god**. Disponível em: <https://www.thecontinent.org/_files/ugd/287178_73f3d2af22614e678f277b631a62e491.pdf>. Acesso em: 11 jun. 2023.

Artificial intelligence and human rights. 1. ed. [s.l.] Dykinson, S.L., 2021.

BENDER, E. M. **The Power of Linguistics - Unpacking Natural Language Processing Ethics with Emily M. Bender.** [**Podcast**]. Disponível em: <https://www.radicalai.org/e16-emily-bender>. Acesso em: 7 abr. 2023.

BENDER, E. M. et al. **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** . Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...: FAccT '21.New York, NY, USA: Association for Computing Machinery, 2021. Disponível em: <https://doi.org/10.1145/3442188.3445922>

BENDER, E. M. **You Are Not a Parrot And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this**. Disponível em: <https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html>. Acesso em: 9 abr. 2023.

BUOLAMWINI, J.; GEBRU, T. **Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification**. (S. A. Friedler, C. Wilson, Eds.)Proceedings of the 1st Conference on Fairness, Accountability and Transparency. **Anais**...: Proceedings

of Machine Learning Research.PMLR, 2018. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>

CAVALIERE, P.; ROMEO, G. From Poisons to Antidotes: Algorithms as Democracy Boosters. **European Journal of Risk Regulation**, v. 13, n. 3, p. 421–442, 2022.

CHANDRAN, R. **Indigenous groups in NZ, US fear colonisation as AI learns their languages**. Disponível em: <https://www.context.news/ai/nz-us-indigenous-fear-colonisation-as-bots-learn-their-languages>. Acesso em: 7 abr. 2023.

COECKELBERGH, M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. **Science and Engineering Ethics**, v. 26, p. 2051–2068, 2020.

COMMISSION, E. **Proposal for a Regulation laying down harmonised rules on artificial intelligence**. Disponível em: < https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence >. Acesso em: 28 ago. 2023.

CORRÊA, N. K. et al. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. **Patterns**, v. 4, n. 10, p. 100857, 2023.

CRUZ, B. S. **Concessionária do Metrô de SP é processada por ter câmeras que leem nossas emoções**. Disponível em: < https://www.uol.com.br/tilt/noticias/redacao/2018/08/31/concessionaria-do-metro-de-sp-e-processada-por-ter-cameras-que-leem-emocoes.htm >. Acesso em: 29 ago. 2023.

CRUZ, B. S. **Racismo Calculado**. Disponível em: < https://www.uol.com.br/tilt/reportagens-especiais/como-os-algoritmos-espalham-racismo/#cover >. Acesso em: 29 ago. 2023.

EMPOLI, G. DA. **Os engenheiros do caos: Como as fake news, as teorias da conspiração e os algoritmos estão sendo utilizados para disseminar ódio, medo e influenciar eleições**. [s.l.] Vestígio Editora, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016. v. 1

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. **Construing Experience Through Meaning: A Language Based Approach to Cognition**. [s.l.] Continuum, 1999.

HEIKKILÄ, M. **Why you shouldn't trust AI search engines**. Disponível em: <https://www.technologyreview.com/2023/02/14/1068498/why-you-shouldnt-trust-ai-search-engines/>. Acesso em: 9 abr. 2023.

HEIKKILÄ, M. **The viral AI avatar app Lensa undressed me—without my consent**. Disponível em: < https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/>. Acesso em: 28 ago. 2023.

HORA, N. DA. **Coded Bias: linguagem acessível para entender vieses em algoritmos**. Disponível em: < https://mittechreview.com.br/coded-bias-linguagem-acessivel-para-entender-vieses-em-algoritmos/>. Acesso em: 7 abr. 2023.

HORA, N. DA. **Ética em IA: a pergunta que não estamos fazendo**. Disponível em: <https://mittechreview.com.br/etica-em-ia-a-pergunta-que-nao-estamos-fazendo/>. Acesso em: 7 abr. 2023.

INFOBASE. **Inteligência Artificial e a perpetuação do racismo**. Disponível em: <https://infobase.com.br/inteligencia-artificial-e-a-perpetuacao-do-racismo/>. Acesso em: 28 ago. 2023.

KANTAYYA, S. **Coded Bias**. Disponível em: < https://www.codedbias.com>. Acesso em: 7 abr. 2023.

LGPD. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 9 abr. 2023.

NUNES, P. **LEVANTAMENTO REVELA QUE 90,5% DOS PRESOS POR MONITORAMENTO FACIAL NO BRASIL SÃO NEGROS**. Disponível em: < https://www.intercept.com.br/2019/11/21/presos-monitoramento-facial-brasil-negros/>. Acesso em: 28 ago. 2023.

O'NEIL, C. **Algoritmos de Destruição em Massa**. [s.l.] Editora Rua do Sabão, 2021.

OECD. **The OECD Framework for the Classification of AI systems**. Disponível em: < https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf>. Acesso em: 28 ago. 2023.

OECD. **OECD Employment Outlook 2023**. [s.l: s.n.]. p. 267

OPENAI. **ChatGPT: OpenA's conversational AI model**. Disponível em: <https://openai.com/blog/chatgpt/>. Acesso em: 7 abr. 2023.

PERRIGO, B. Disponível em: <https://time.com/6247678/openai-chatgpt-kenya-workers/>. Acesso em: 9 abr. 2023.

REVIEW, M. T. **Um aplicativo de Inteligência Artificial que "desnudava" mulheres mostra como as deepfakes prejudicam os mais vulneráveis**. Disponível em: < https://mittechreview.com.br/um-aplicativo-de-inteligencia-artificial-que-desnudava-mulheres-mostra-como-as-deepfakes-prejudicam-os-mais-vulneraveis/>. Acesso em: 28 ago. 2023.

RUSSEL, S. **Human Compatible Artificial Intelligence and the Problem of Control**. [s.l.] Penguin Books, 2019.

SARTORI, L.; THEODOROU, A. A Sociotechnical Perspective for the Future of AI:

Narratives, Inequalities, and Human Control. **Ethics and Inf. Technol.**, v. 24, n. 1, mar. 2022.

SMITH, G.; RUSTAGI, I. **Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook**. [s.l.] Berkeley Haas Center for Equity, Gender; Leadership, 2020.

UNESCO. **Beijing consensus on artificial intelligence and education**. UNESCO Paris, 2019.

UNESCO, D. G. **Recomendação sobre a Ética da Inteligência Artificial**. Disponível em: < https://unesdoc.unesco.org/ark:/48223/pf0000381137_por >. Acesso em: 28 ago. 2023.

UNICEF. **Declaração Universal dos Direitos Humanos**. Disponível em: < https://www.unicef.org/brazil/declaracao-universal-dos-direitos-humanos>. Acesso em: 28 ago. 2023.