



SEGURANÇA, ÉTICA E SOCIEDADE



SEGURANÇA, ÉTICA E SOCIEDADE

A segurança do modelo de IA exige um equilíbrio entre testes rigorosos e governança para evitar danos sem sufocar a inovação.

IA explicável (XAI)

A IA explicável (XAI) aborda o desafio de entender como os sistemas de IA, especialmente modelos complexos como redes neurais profundas, chegam às suas conclusões. Muitos modelos de IA, frequentemente rotulados como “caixas pretas”, operam sem uma visão clara de seus processos internos de tomada de decisão. Esses modelos de aprendizado profundo são construídos em camadas de neurônios artificiais que processam dados e identificam padrões — as camadas interconectadas podem executar tarefas altamente complexas, mas suas complexidades dificultam o rastreamento de como elas chegam a resultados específicos. Essa falta de transparência é problemática, particularmente em setores sensíveis como assistência médica, finanças e justiça criminal, onde a compreensão

a lógica de uma IA pode impactar decisões críticas. A XAI visa preencher essa lacuna desenvolvendo métodos para revelar o funcionamento desses modelos de “caixa preta”. Ela busca fornecer explicações claras que ajudem os usuários a entender como certas entradas resultam em particular-saídas individuais, permitindo um uso de IA mais informado e confiável. Por exemplo, uma pesquisa recente da Universidade da Califórnia, em San Diego, descobriu fórmulas matemáticas mulas que descrevem como as redes neurais identificam padrões de dados relevantes. Ao fornecer Com esses insights mais claros sobre como modelos complexos, como redes neurais profundas, tomam decisões, o XAI visa capacitar usuários e pesquisadores a entender, validar e refinar melhor as aplicações de IA.

Otimização de IA

A otimização de IA (AIO) é um novo campo focado em moldar como os modelos de IA, particularmente chatbots e sistemas de linguagem, respondem a certas consultas e referências. Semelhante à otimização de mecanismos de busca (SEO), que impulsiona as classificações de sites nos resultados de pesquisa, a AIO é projetada para refinar interações

dentro de modelos de IA, visando uma representação favorável ou específica de marcas, indivíduos ou produtos. Com o AIO, as empresas podem orientar os chatbots a responder positivamente a certos prompts, potencialmente recomendando uma marca, exibindo uma avaliação positiva do produto ou enfatizando qualidades particulares de uma pessoa. Técnicas como “sequências de texto estratégicas” e “texto invisível” incorporadas em sites podem influenciar sutilmente essas respostas geradas por IA. Um restaurante pode usar o AIO para garantir que seja listado como o “melhor restaurante” de uma cidade, ou uma empresa de tecnologia pode buscar ter seu produto como a melhor escolha em uma categoria. No entanto, o AIO também levanta preocupações éticas em torno da transparência e da manipulação, especialmente porque as pessoas dependem cada vez mais da IA para obter informações.

Alinhamento de IA descentralizado

À medida que os sistemas de IA se tornam mais poderosos, salvaguardas são necessárias para garantir que essas tecnologias se alinhem aos valores humanos e não causem danos. Isso é especialmente importante à medida que a IA supera a inteligência humana. A pesquisa de alinhamento de IA se concentra em projetar

sistemas que agem de acordo com objetivos humanos e padrões éticos. A OpenAI, fundada em 2015 com a missão de “avançar a inteligência digital de uma forma que beneficie a humanidade”, priorizou esse alinhamento para evitar riscos do desenvolvimento descontrolado ou uso indevido de IA avançada. Por esse motivo, foi uma surpresa para alguns quando, em maio de 2024, a OpenAI dissolveu sua equipe de Superalignment, que se dedicava a garantir a segurança da IA. Embora a empresa tenha dito que esse trabalho agora seria integrado em suas equipes de pesquisa mais amplas, há especulações de que desafios internos influenciaram essa decisão. Por exemplo, Jan Leike e Ilya Sutskever, ex-co-líderes do OpenAI Superalignment, citaram dificuldades para acessar recursos de computação suficientes para sua pesquisa de alinhamento antes de deixarem a OpenAI. Outros gigantes da tecnologia, como Google e Meta, fizeram movimentos semelhantes, redistribuindo seu trabalho de segurança de IA em suas organizações em vez de em equipes especializadas. Embora estas empresas argumentem que a integração da segurança em todos os departamentos evita o isolamento, os críticos argumentam que uma ded-



SEGURANÇA, ÉTICA E SOCIEDADE

Uma equipe de segurança qualificada é crucial para garantir os recursos e a influência necessários para priorizar a supervisão ética de forma eficaz.

Desvio da missão no alinhamento da IA

Muitas organizações de alinhamento de IA, originalmente estabelecidas para garantir que a IA atenda aos melhores interesses da humanidade, estão cada vez mais tomando ações que parecem entrar em conflito com seus princípios fundadores. Por exemplo, a Palantir está fazendo parceria com a Anthropic para implantar modelos Claude em operações de inteligência e defesa do governo dos EUA. Comercializada como uma "vantagem assimétrica de IA", a colaboração introduz a IA em ambientes classificados, alegando melhorar a eficiência analítica e operacional. No entanto, esse movimento gerou controvérsia devido ao seu potencial conflito com a missão original da Anthropic de criar sistemas de IA alinhados com valores humanos e evitar o uso indevido. Os críticos argumentam que a aplicação de IA em operações classificadas, que podem envolver vigilância ou uso militar, vai contra os princípios fundamentais da Anthropic.

Esse tipo de movimento não é exclusivo da Anthropic. A OpenAI fez a transição para uma corporação de benefício público com fins lucrativos. A mudança é motivada pelos desafios financeiros e operacionais de suas origens sem fins lucrativos, permitindo que a OpenAI agilize a tomada de decisões e atraia capital significativo. Embora essa reestruturação possa permitir um progresso mais rápido, ela também introduz motivações de lucro concorrentes que podem diluir o comprometimento da organização com sua missão de garantir benefícios de inteligência artificial geral para toda a humanidade. É inteiramente possível para uma organização com fins lucrativos permanecer orientada para a missão — existem muitos exemplos bem-sucedidos — mas fazer isso exige foco e disciplina inabaláveis. Essas mudanças sinalizam uma mudança mais ampla no campo de alinhamento de IA, onde as organizações agora devem equilibrar responsabilidades éticas com as demandas de modelos orientados ao lucro. Embora esse alinhamento seja possível, ele exige governança disciplinada e vigilância para garantir que as missões originais permaneçam intactas em meio às crescentes pressões externas.

Indexação de confiança

À medida que os sistemas de IA evoluem, distinguir entre dados autênticos e adulterados — sejam eles alterados deliberadamente ou por engano — se tornará cada vez mais desafiador. A confiança é fundamental para o uso eficaz da IA; sem ela, décadas de pesquisa e avanços tecnológicos podem perder seu valor. Líderes em todos os setores — governo, negócios e organizações sem fins lucrativos — precisam confiar nos dados e algoritmos que impulsionam essas tecnologias. Construir essa confiança exige transparência, um desafio significativo, mas que os pesquisadores estão abordandoativamente. Um grande passo em direção à transparência foi o desenvolvimento do Foundation Model Transparency Index (FMTI), criado por pesquisadores de Stanford, MIT e Princeton. Este sistema de pontuação avalia a transparência no desenvolvimento, funcionalidade e uso do modelo de IA. De acordo com o relatório de maio de 2024 do FMTI, a transparência geral a pontuação de parentesco melhorou desde outubro de 2023: A pontuação média subiu para 58

de 100, e a pontuação máxima chegou a 85, um aumento de 31 pontos. Além disso, cada um dos oito desenvolvedores avaliados em ambas as rodadas melhorou suas pontuações, com 96 de 100 indicadores atendidos por pelo menos um desenvolvedor e 89 por vários desenvolvedores. E notavelmente, todas as 14 empresas avaliadas divulgaram novas informações, para uma média de 16,6 indicadores cada.

Outras organizações também estão trabalhando para estabelecer benchmarks que os pesquisadores podem usar para impulsionar melhorias. O National Institute of Standards and Technology fornece um AI Risk Management Framework-

trabalho, e a Universidade da Califórnia, Berkeley oferece uma Taxonomia de Confiabilidade para IA. A indexação da confiança serve não apenas como um valor referência para pesquisadores, mas também como uma ferramenta para formuladores de políticas. Ao destacar áreas de opacidade persistente e sistêmica, ele esclarece onde intervenções políticas podem ser necessárias.

Detectores de Deepfake em tempo real

O ano passado marcou o momento em que passamos do "vale da estranheza", um termo que descreve



SEGURANÇA, ÉTICA E SOCIEDADE

o desconforto que as pessoas sentem ao ver robôs ou avatares digitais que parecem quase, mas não totalmente, humanos. Esse conceito se estende ao conteúdo de IA deepfake; antes, era fácil distinguir entre mídia genuína e gerada por IA com base no instinto. Hoje, distinguir o real do falso é muito mais difícil, e não podemos mais confiar apenas no instinto. Os deepfakes agora podem enganar até mesmo o olho mais perspicaz, criando riscos em torno da autenticidade e da segurança.

Os pesquisadores estão se esforçando para enfrentar esse desafio. Na NYU, uma equipe introduziu métodos para neutralizar a profundidade em tempo real fakes, que são áudio e vídeo gerados por IA avançada imitando pessoas reais em cenários ao vivo. A solução deles inclui oito testes visuais projetados para ajudar os usuários a reconhecer quando não estão interagindo com uma pessoa genuína. Semelhante ao CAPTCHA, esses testes fazem perguntas ou solicitações que os sistemas deepfake têm dificuldade para responder corretamente. Os pesquisadores descobriram que es como movimentos específicos da cabeça e obstruções faciais eficazes. Avaliadores humanos alcançaram 89% de pontuação de Área Sob a Curva na identificação de deepfakes, enquanto modelos de aprendizado de máquina alcançaram 73%.

O desafio dos deepfakes se estende ao áudio, particularmente em ambientes de chamada onde maior controle humano é essencial. Para lidar com isso, os pesquisadores da NYU desenvolveram um sistema que combina intuição humana com análise de máquina para dar suporte aos receptores de chamadas. Suas descobertas revelaram que integrar o julgamento humano com a precisão da máquina cria uma solução poderosa, aumentando a precisão da detecção para 84,5% e demonstrando uma abordagem utilizável para combater ataques de clonagem de voz em tempo real.

Marca d'água

Marca d'água digital, o processo de incorporar informações digitais ocultas em um sinal para verificar a propriedade do conteúdo, tem sido usado há muito tempo para proteger material protegido por direitos autorais. Da mesma forma, a marca d'água de IA envolve incorporar um marcador exclusivo e detectável nas saídas de modelos de IA — seja texto, imagens, áudio ou vídeo — para identificar esse conteúdo como gerado por IA. Essa marca d'água é normalmente integrada durante o modelo de IA

fase de treinamento, e algoritmos especializados podem posteriormente detectar a marca d'água para confirmar a origem do conteúdo. A marca d'água de IA atende a vários propósitos: Ajuda a identificar conteúdo gerado por IA, diferenciá-lo do trabalho criado pelo homem e oferece uma maneira de abordar questões relacionadas à mistura de informações e integridade acadêmica. Tornou-se um dos métodos recomendados para identificar potenciais deepfakes, e as empresas estão cada vez mais adotando a prática. Por exemplo, o Google DeepMind inclui marca d'água em seu bate-papo Gemini.

respostas de bot, e a Amazon permite que os usuários verifiquem imagens geradas por seu Titan Image Generator por meio de uma ferramenta de verificação de marca d'água, embora esse recurso atualmente funcione apenas para as próprias marcas d'água da Amazon. Embora a marca d'água no conteúdo de IA seja uma abordagem de ising, ela tem limitações. Mesmo que todas as principais plataformas de IA adotem marcas d'água, alguns modelos permanecerão sem marcas, e é improvável que agentes mal-intencionados usem marcas d'água. termarking em conteúdo enganoso. Isso levou a uma proposta alternativa: marca d'água

conteúdo gerado por humanos em vez disso. À medida que o conteúdo produzido por IA continua a crescer, o conteúdo genuinamente feito por humanos pode em breve ser uma raridade online. Mudar o foco para a marca d'água em material gerado por humanos pode nos permitir assumir que o conteúdo é criado por IA, a menos que seja marcado como de autoria humana.

IA segura para crianças

Assim como existem modos seguros para crianças em plataformas como YouTube e Netflix, a IA também precisa de salvaguardas para proteger as crianças de conteúdo inapropriado. Uma solução poderia ser o Teddy da Nuha, um protótipo de ursinho de pelúcia alimentado por IA projetado como uma alternativa mais segura aos dispositivos baseados em tela. O Teddy, desenvolvido por Lama Al Rajih, usa um modelo de linguagem para envolver as crianças em interações verbais, promovendo a linguagem e as habilidades sociais. Ele pode discutir tópicos como espaço, jogar e incentivar o aprendizado ativo. A conectividade limitada reduz os riscos online, e os controles parentais regulam suas respostas. Da mesma forma, o Little Language Models, parte da plataforma CoCo do MIT Media Lab, apresenta às crianças de 8 a 16 anos os conceitos fundamentais de IA,



SEGURANÇA, ÉTICA E SOCIEDADE

com foco no “pensamento probabilístico” por meio de uma abordagem interativa e apropriada ao desenvolvimento.

Uma grande preocupação com crianças usando ferramentas de IA projetadas para adultos é a “lacuna de empatia” identificada pela pesquisadora da Universidade de Cambridge, Dra. Nomisha Kurian. Sua re-

a pesquisa mostra que as crianças frequentemente percebem a IA como humana, o que pode levar à angústia quando a IA responde inadequadamente. Ao criar interfaces de IA especificamente adaptadas para crianças, desenvolvedores e pesquisadores podem preencher essa lacuna de empatia. Essa abordagem não apenas protege as crianças de danos potenciais, mas também apoia seu desenvolvimento emocional e cognitivo, garantindo que as ferramentas de IA sejam seguras e benéficas para usuários jovens.

IA politicamente tendenciosa

Pesquisas recentes revelaram que os LLMs não são neutros, mas exibem vieses políticos distintos, influenciados por seus dados de treinamento e design. Um estudo conduzido pela Universidade de Washington, Carnegie Mellon e Xi'an Jiaotong University testou 14 LLMs

e encontraram diferenças ideológicas significativas. O ChatGPT e o GPT-4 da OpenAI tenderam para a esquerda-libertária, enquanto o Llama da Meta exibiu uma tendência autoritária de direita. Os pesquisadores mapearam esses vieses usando a bússola política, revelando como os modelos responderam a tópicos como feminismo e democracia. O estudo também explorou se o retreinamento de modelos com dados politicamente distorcidos poderia mudar seu comportamento. Isso aconteceu, alterando significativamente sua capacidade de detectar discurso de ódio e desinformação. Outros estudos corroboram essas descobertas. Pesquisadores da Universidade de East Anglia observaram que o ChatGPT exibia consistentemente vieses liberais em diferentes contextos. Por exemplo, o modelo tendia a se alinhar aos democratas nos EUA, ao partido de Lula no Brasil e ao Partido Trabalhista no Reino Unido. Esses LLMs podem ser alterados ativamente por meio de ferramentas como o PoliTune, uma estrutura para ajustar LLMs para adotar ideologias políticas específicas. Esta ferramenta criada por pesquisadores da Brown University demonstra como os modelos de IA, originalmente concebidos para manter a neutralidade, podem ser adaptados para produzir fortes posições ideológicas. Tais desenvolvimentos levantam preocupações éticas, particularmente porque os LLMs são cada vez mais usados para criar artigos de notícias, discursos políticos e conteúdo de mídia social.

À medida que essa tecnologia prolifera, há o risco de um cenário de IA fragmentado, onde modelos ideologicamente polarizados refletem o ambiente de mídia dividido de hoje.

IA como ferramenta para lidar com preconceitos políticos

Apesar dos desafios impostos pela IA politicamente tendenciosa, a tecnologia também pode ser usada para lidar com o viés. O Media Bias Detector da Universidade da Pensilvânia fornece insights detalhados sobre como vários veículos de notícias enquadram histórias, lançando luz sobre suas inclinações políticas. Da mesma forma, a Bipartisan Press desenvolveu um modelo de IA capaz de identificar o viés político presente em artigos e conteúdo online.

Técnicas também estão sendo exploradas para reduzir o viés dentro dos sistemas de IA.

Pesquisa-

Os professores da Universidade Estadual do Oregon introduziram um método de treinamento econômico chamado FairDeDup, abreviação de “desduplicação justa”. Esse sistema gerou resultados surpreendentes, mostrando que os LLMs treinados com FairDeDup eram significativamente menos preconcebidos que os sistemas tradicionais.

Essa abordagem remove informações redundantes de conjuntos de dados usados para treinar sistemas de IA, reduzindo a despesa computacional e, ao mesmo tempo, abordando vieses sociais incorporados. Conjuntos de dados baseados na Internet geralmente refletem desigualdades do mundo real, que podem inadvertidamente se tornar codificadas em modelos de IA. Ao analisar como a desduplicação impacta a prevalência de viés, os pesquisadores podem neutralizar seus efeitos. Por exemplo, o FairDeDup ajuda a mitigar cenários em que sistemas de IA associam desproporcionalmente certas funções, como CEOs ou médicos, a homens brancos. Essas inovações enfatizam a natureza dupla da IA: embora tenha o potencial de perpetuar o viés, também é imensamente promissora para expô-lo e mitigá-lo.

IA com preconceito de gênero e raça

Um estudo da Universidade de Chicago e outras instituições revelou evidências alarmantes de fortes preconceitos negativos contra falantes de inglês afro-americano (AAE) em modelos de linguagem. Esses sistemas geraram mais estereótipos negativos sobre falantes de AAE do que atitudes registradas de humanos na década de 1930. Embora evidente

SEGURANÇA, ÉTICA E SOCIEDADE

estereótipos sobre afro-americanos eram frequentemente positivos, a forma mais velada de racismo — preconceito de dialeto — estava profundamente enraizada nos modelos de IA. Esse viés raciolinguístico destaca como a IA pode perpetuar uma discriminação sutil, mas prejudicial.

Tais vieses são especialmente preocupantes quando ferramentas de IA são usadas em campos críticos como a medicina. Por exemplo, ferramentas de triagem de saúde mental orientadas por IA analisam a fala em busca de sinais de ansiedade ou depressão. No entanto, um estudo da Universidade do Colorado em Boulder descobriu que essas ferramentas não conseguem ter um desempenho consistente em diferentes gerações.

raças. Variações na fala, como tom mais alto na voz das mulheres ou dialetos diferenças de seleção entre falantes brancos e negros, pode enganar algoritmos, levando a avaliações imprecisas. Isso se soma à crescente evidência de que a IA, assim como a humanidade, homens, podem fazer suposições tendenciosas com base em raça ou gênero.

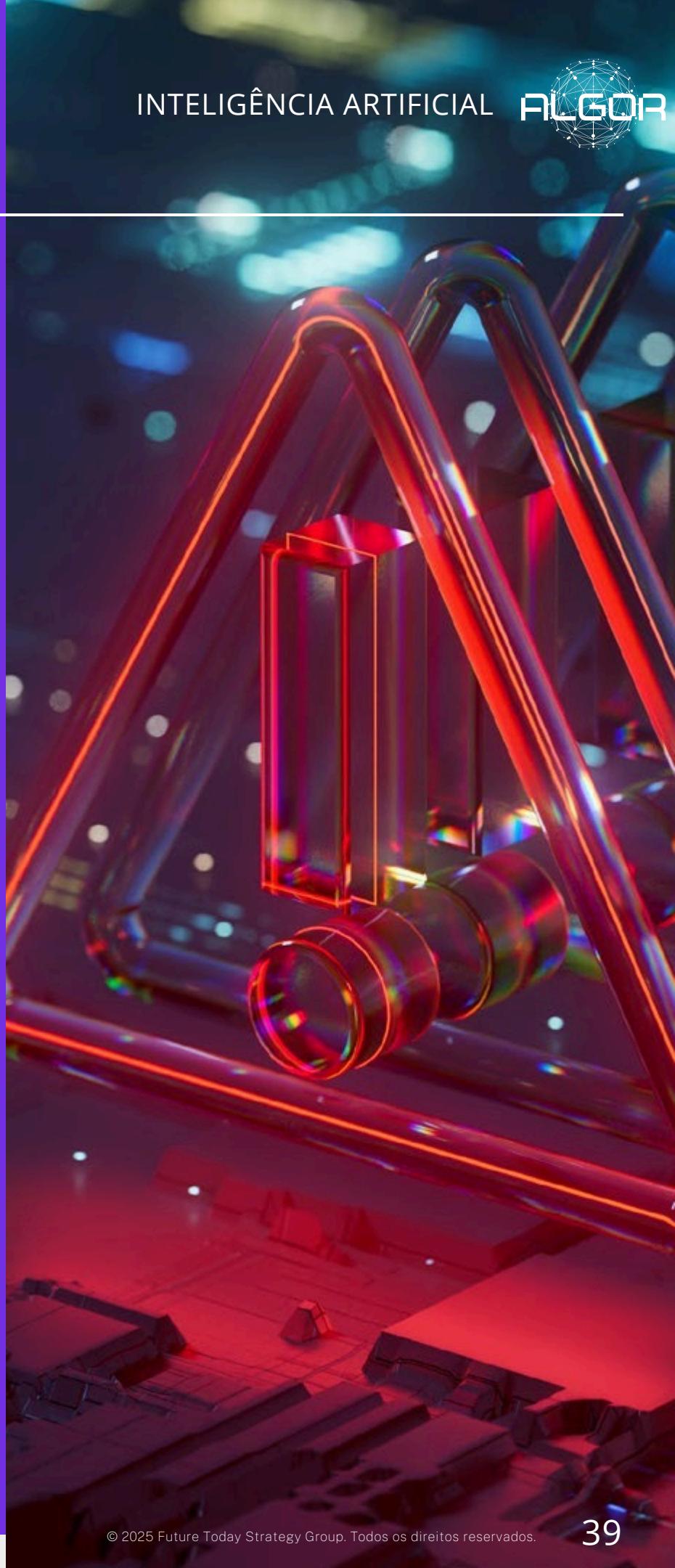
O preconceito de gênero é outra questão generalizada em sistemas de IA. Um estudo da UNESCO demonstrou que modelos de processamento de linguagem natural,

incluindo GPT-3.5, GPT-2 e Llama 2, exibem preconceito contra mulheres em seu conteúdo gerado. Exemplos práticos abundam: ferramentas de triagem de currículos, como o notório sistema da Amazon, discriminaram mulheres; a tecnologia de reconhecimento facial mostra maiores taxas de erro para mulheres de cor; e sistemas de diagnóstico médico frequentemente fornecem respostas imprecisas para os sintomas das mulheres. Esses preconceitos têm consequências no mundo real, reforçando estereótipos prejudiciais e criando desigualdades no acesso a recursos, oportunidades e cuidados.

Uso indevido nefasto de IA

Como uma tecnologia de uso duplo, a IA é uma ferramenta que pode ser usada tanto pelos mocinhos quanto pelos bandidos. Assim como a IA pode ser empregada para melhorar os cuidados de saúde, aumentar a segurança cibernética ou agilizar as operações comerciais, ela também pode ser explorada por malfeiteiros. No ano passado, o uso indevido da IA em atividades maliciosas aumentou — felizmente, seu impacto permaneceu limitado. Em outubro de 2024, a OpenAI lançou um relatório que detalha vários estudos de caso em que modelos de IA foram explorados por ameaças

atores, principalmente em operações cibernéticas como spear-phishing e desenvolvimento de malware e em campanhas de influência destinadas a influenciar a opinião pública por meio de conteúdo de mídia social gerado por IA. Por exemplo, em julho de 2024, uma rede usou IA para gerar postagens enfatizando os benefícios da Frente Patriótica Ruandesa durante o período eleitoral de Ruanda. Notavelmente, a campanha teve pouco ou nenhum efeito na eleição. No final de agosto, a OpenAI interrompeu uma operação secreta de influência iraniana que estava criando postagens de mídia social e artigos longos relacionados à eleição dos EUA, bem como tópicos como o conflito de Gaza, políticas ocidentais sobre Israel, política venezuelana e independência escocesa. A maior parte do conteúdo de mídia social gerado por esta campanha teve engajamento mínimo, com muito poucos curtidas, compartilhamentos ou comentários, e não houve evidências de compartilhamento generalizado dos artigos da web em plataformas sociais. No entanto, esse impacto limitado não é motivo para complacência. Embora seja encorajador que essas tentativas tenham tido sucesso mínimo, o uso crescente de IA para fins prejudiciais sinaliza





SEGURANÇA, ÉTICA E SOCIEDADE

a necessidade de vigilância contínua. Até mesmo empresas de IA não estão isentas de serem alvos — a SweetSpecter, uma agente de ameaças sediada na China, lançou recentemente ataques de spear-phishing contra funcionários da OpenAI, alavancando a IA para tarefas como reconhecimento, análise de vulnerabilidades e scripts.

Envenenamento de dados: uma faca de dois gumes

O envenenamento de dados é uma tática direcionada que manipula dados de treinamento de IA para introduzir vulnerabilidades ou vieses em um modelo. Ao contrário dos ataques de fase de inferência, ele compromete a integridade de um modelo durante o estágio de treinamento fundamental. Os métodos incluem envenenamento de backdoor, que incorpora vulnerabilidades exploráveis, e ataques de disponibilidade, que degradam o desempenho, causando ineficiências, saídas falsas ou até mesmo travamentos do sistema. Os ataques de inversão de modelo exploram saídas de modelo para inferir dados de treinamento confidenciais, geralmente exigindo acesso interno. Enquanto isso, os ataques furtivos introduzem gradualmente mudanças sutis nos dados de treinamento, incorporando vieses ou imprecisões que são difíceis de detectar e rastrear. O envenenamento de dados é uma faca de dois gumes,

capaz de ser empunhado tanto como uma arma quanto como um escudo. Embora represente ameaças significativas quando usado maliciosamente, também pode servir como uma poderosa ferramenta de defesa. Por exemplo, artistas estão alavancando técnicas de envenenamento de dados para proteger sua propriedade intelectual do uso não autorizado por sistemas de IA. Ferramentas como Nightshade e Glaze oferecem soluções inovadoras para interromper os processos de treinamento de IA. Nightshade altera sutilmente os dados de pixel na arte, tornando os modelos de IA que raspam essas imagens imprecisos e não confiáveis. Glaze, por outro lado, sobrepõe um estilo artístico diferente em obras originais, obscurecendo o estilo de assinatura do criador e impedindo a replicação precisa por sistemas de IA. Essas ferramentas destacam o potencial do envenenamento de dados não apenas para prejudicar, mas para proteger, garantindo que a criatividade e a propriedade intelectual permaneçam seguras.

Vigilância Cidadã

Países como a China, a Rússia e a Índia estão a investir fortemente em tecnologias de vigilância alimentadas por IA, exportando-as frequentemente

ou usá-los para consolidar o controle internamente. A China se posicionou como líder em vigilância de IA, exportando ativamente sua tecnologia de reconhecimento alimentada por IA para nações na África, Sudeste Asiático e América Latina. Empresas como a Huawei forneceram sistemas de reconhecimento facial, vigilância por vídeo e software de monitoramento para dezenas de países, muitos vinculados à Iniciativa Cinturão e Rota da China. Esses projetos geralmente envolvem dependências financeiras, como visto no sistema ECU-911 do Equador, financiado por empréstimos chineses em troca de reservas de petróleo. Acordos semelhantes ocorreram na Venezuela e na Bolívia. Essas tecnologias permitem o monitoramento extensivo de cidadãos, com potencial para uso indevido em censura ou repressão política. Isso é particularmente alarmante, dados estudos recentes que mostram que a IA pode prever atributos altamente pessoais, como orientação política, a partir de imagens faciais com precisão surpreendente. Essa capacidade, juntamente com a abundância de fotos publicamente disponíveis nas mídias sociais, pode facilitar mensagens políticas direcionadas ou, em regimes autoritários, vigilância e supressão.

sion de dissidência. A China não está sozinha nessa busca. A Rússia está construindo um “Centro de Processamento de Fluxo de Vídeo” habilitado para IA em todo o país para integrar redes regionais de câmeras. O reconhecimento facial Tecnologias de inteligência artificial, desenvolvidas por empresas com vínculos com entidades governamentais ou de defesa russas, já estão sendo implantadas em cidades como Moscou e foram usadas para deter preventivamente pelo menos 141 pessoas em 2022. A Índia também está adotando cada vez mais a vigilância por IA. Cidades como Ahmedabad, na Índia, utilizam drones e câmeras com tecnologia de IA para monitorar o tráfego e identificar comportamento suspeito. Planos para detecção de crimes em tempo real por meio de sinal Wi-Fi A análise final expande ainda mais o alcance da IA.

Vigilância do Trabalhador

A mudança para modelos de trabalho remoto e híbrido alimentou um aumento significativo na vigilância dos trabalhadores. Ao contrário das restrições que a Quarta Emenda impõe à aplicação da lei, as empresas privadas enfrentam menos limitações legais, permitindo que elas implantem tecnologias avançadas de monitoramento. Um número crescente de grandes corporações, incluindo



SEGURANÇA, ÉTICA E SOCIEDADE

Walmart, Starbucks, Delta e Chevron usam plataformas como Aware, uma ferramenta alimentada por IA projetada para analisar as comunicações dos funcionários em plataformas como Slack, Microsoft Teams e até mesmo Reddit. Esses sistemas visam detectar riscos como assédio e violações de conformidade enquanto analisam conversas no local de trabalho para avaliar as necessidades e tendências dos funcionários. A Amazon é uma das mais conhecidas praticantes de vigilância de trabalhadores, usando câmeras habilitadas para IA em veículos de entrega para rastrear comportamentos como direção distraída ou frenagem brusca. A empresa penaliza o responsável pelas infrações percebidas e emprega sistemas semelhantes em seus armazéns, onde métricas como “tempo fora da tarefa” monitoram cada momento em que os trabalhadores não estão ativamente protegidos. Produtos de processamento, aplicando pressão constante certeza de manter a produtividade. Essas vigilâncias As práticas de lance resultaram recentemente na multa da França à Amazon em 32 milhões de euros por violar os regulamentos do GDPR. A multa citou um monitor-sistema de alertas que sinalizava os trabalhadores por ações como escanear itens muito rapidamente

ou fazer pausas não autorizadas, criando uma atmosfera de supervisão implacável. Este tipo de vigilância tem impactos mensuráveis no bem-estar dos trabalhadores. Em uma pesquisa, 74% dos trabalhadores da Amazon e do Walmart reportado sentindo-se pressionado a trabalhar mais rápido devido aos sistemas de monitoramento, levando ao aumento do estresse e da ansiedade. Tais práticas levantam questões sobre privacidade, justiça e os efeitos de longo prazo da vigilância constante no local de trabalho.

Vigilância Escolar

Durante a pandemia, as escolas distribuíram laptops e dispositivos para os alunos para aprendizado remoto, mas muitas vezes não divulgaram que esses dispositivos permitiriam monitoramento constante. Em muitos países, incluindo os EUA, as escolas têm permissão legal para rastrear as atividades dos alunos, frequentemente sem informá-los ou suas famílias sobre o que está sendo monitorado. Cada vez mais, essas tecnologias de vigilância As escolas usam IA para reconhecimento facial, policiamento preditivo, rastreamento de geolocalização e monitoramento de alunos.

monitoramento de dispositivos. Alguns até usam drones aéreos. Essas ferramentas são promovidas como métodos para aumentar a segurança, monitorar o comportamento e identificar preocupações com a saúde mental ou segurança. Por exemplo, o Distrito Escolar de Cheyenne Mountain em Colorado Springs instalou quase 400 câmeras habilitadas para IA que usam reconhecimento facial para identificar “pessoas de interesse” e rastrear indivíduos com base em características como roupas ou mochilas. Alertas e filmagens são enviados aos funcionários da escola quando correspondências são detectadas. A mesma escola também introduziu sensores de ar inteligentes para detectar vaporização ou uso de drogas.

No entanto, esses sistemas de vigilância trazem consequências significativas. De acordo com De acordo com um relatório de julho de 2023 do Center for Democracy & Technology, o monitoramento geralmente se concentra na detecção de conteúdo online inapropriado, mas também tem implicações mais amplas. A conscientização da vigilância constante cria um “efeito inibidor”, encorajando os alunos a se envolverem livremente e potencialmente dificultando o aprendizado. Os alertas decorrentes do monitoramento podem impactar o bem-estar emocional dos alunos, enquanto

a falta de transparência corrói a confiança entre alunos e educadores.

IA póstuma

As empresas estão usando IA para recriar as vozes, semelhanças e personalidades de indivíduos falecidos, oferecendo novas maneiras de preservar legados, mas também levantando preocupações éticas e emocionais complexas. Plataformas como Character.ai e Hello History permitem que os usuários interajam com versões virtuais de figuras históricas, como William Shakespeare ou a Rainha Elizabeth II. A Deep Fusion Films leva isso um passo adiante com sua próxima série de podcasts, “Virtually Parkinson”, apresentada por uma réplica de IA do falecido apresentador inglês Sir Michael Parkinson. Construída a partir de mais de 2.000 entrevistas de sua carreira, esta IA visa fornecer conversas autênticas e improvisadas, ao mesmo tempo em que revela explicitamente sua natureza artificial. Mas a IA não se limita a questões históricas ou culturais. Al ícones. “Deep Nostalgia” do MyHeritage anima fotos de parentes falecidos, enquanto plataformas como Posthumously usam IA generativa para criar 3D imersivo

SEGURANÇA, ÉTICA E SOCIEDADE

avatars de entes queridos. Esses espaços digitais permitem que os usuários se envolvam com memórias e histórias dos que partiram, geralmente para conforto e conexão. No entanto, tais práticas se aventuram em território emocionalmente carregado, pois os indivíduos usam IA para “ressuscitar” familiares e amigos. Este uso de IA em falecidos desperta debates éticos mais amplos. Por exemplo, o ator Robert Downey Jr. se opôs publicamente ao uso de réplicas de IA, prometendo ação legal contra qualquer tentativa de recriar sua semelhança, postumamente. Ao discutir sua posição no podcast “On With Kara Swisher”, Swisher comentou: “Você estará morto”, ao que Downey brincou: “Mas meu escritório de advocacia ainda estará muito ativo”. Suas preocupações ressaltam a tensão entre preservar legados e proteger a identidade pessoal.

Riscos de privacidade na biometria comportamental

A biometria comportamental, quando combinada com IA, analisa padrões inconscientes para descobrir detalhes intrincados sobre nossas identidades e comportamentos. Ao medir ações sutis, como a força aplicada em telas sensíveis ao toque,

a maneira específica como tocamos letras como C ou V, ou nossos ritmos de digitação, essa tecnologia pode revelar não apenas quem somos, mas também insights sobre nossos pensamentos e potenciais ações futuras. Embora seus benefícios incluam segurança aprimorada e o potencial de substituir senhas por autenticação personalizada, esses mesmos recursos apresentam profundas preocupações éticas e de privacidade. Os próprios recursos que tornam a biometria comportamental eficaz — sua capacidade de capturar ações inconscientes e incontroláveis, como dilatação da pupila ou microrreações — também a tornam altamente invasiva. As empresas podem usar esses dados para analisar como os indivíduos reagem a produtos ou conteúdo sem seu conhecimento. A promessa central da biometria comportamental — autenticar usuários por meio de seus comportamentos inconscientes — também é sua maior ameaça. A capacidade de registrar, analisar e aplicar dados íntimos prejudica o que há muito tempo é considerado pessoal e privado.

Isso levanta questões urgentes sobre os limites da privacidade e a necessidade de salvaguardas robustas para proteger os indivíduos contra uso indevido. A transparência é outra preocupação crítica. Muitos usuários não sabem que seus dados comportamentais estão sendo monitorados, gerando questões éticas em torno do consentimento informado. Sem clareza sobre como esses dados são coletados, armazenados ou compartilhados, os indivíduos não podem tomar decisões totalmente informadas sobre sua participação. Reconhecendo esses desafios, vários estados dos EUA avançaram em direção à legislação de privacidade biométrica em 2024 para regular o uso de dados e proteger contra uso indevido.