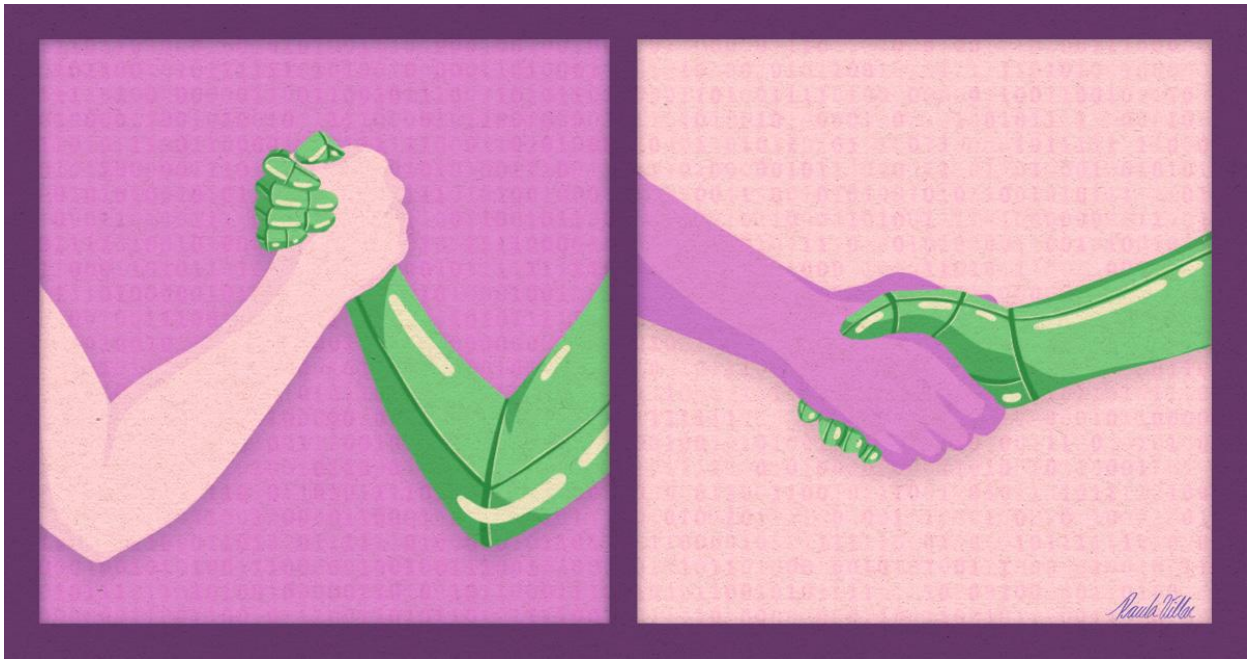

# Inteligências Artificiais entram em campo contra (e a favor) da desinformação

*Na última reportagem da série do Jornal da USP, especialistas alertam para os riscos da IA, incluindo a disseminação de informações falsas e o perigoso uso de 'deepfakes'*

**Texto:** Denis Pacheco*
**Ilustração:** Paula Villar
**Publicado:** 10 de novembro, 2023



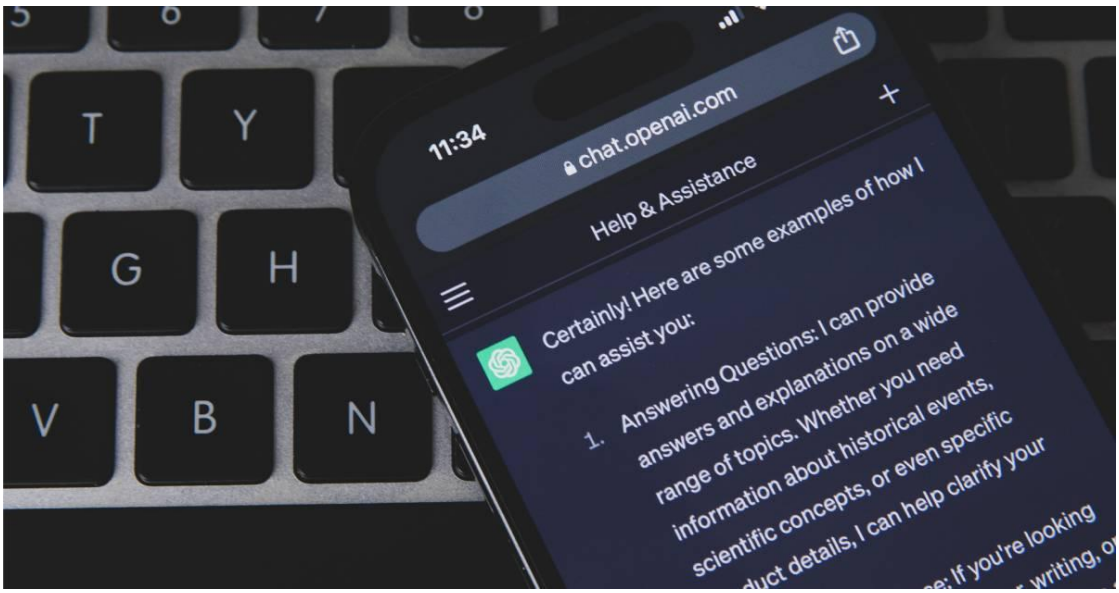
Foi em meados de março deste ano que o professor Fernando Osório, do Instituto de Ciências Matemáticas e de Computação (ICMC) da USP, foi surpreendido com a fala de uma colega de profissão: “O ChatGPT me

matou!”

Curioso, Osório indagou em que circunstâncias ela supostamente teria morrido. “De acordo com o ChatGPT, ela morreu em um acidente de carro, em uma estrada a caminho da universidade onde ela trabalha, em Santa Maria, no Rio Grande do Sul. Ela é de fato professora da PUC no Rio Grande do Sul, mas não existe essa estrada mencionada pelo ChatGPT”, lembra ele ao reforçar que a professora em questão estava não apenas viva e bem, como compartilhava a situação surreal com todos os colegas.

Desde o fatídico - e letal - engano, o professor do Departamento de Sistemas de Computação do ICMC decidiu embarcar em uma jornada nada heróica de transformar o modelo de linguagem, cada vez mais popular, em um “serial killer”.

Pouco mais tarde, naquele mesmo mês, ao solicitar ao modelo de linguagem pequenas biografias de colegas, “eu já estava ‘matando’ pessoas com ele. Eu assassinei o diretor do meu Instituto, o coordenador do Centro de Inteligência Artificial e vários outros colegas”, conta ele. Bem humorado, o professor seguiu para uma próxima etapa em seu projeto funesto: convencer o ChatGPT a criar um texto falando sobre sua própria morte.



*Aparelho com acesso ao ChatGPT - Foto: Reprodução/Jernej Furman /Wikimedia Commons*

A princípio, não foi um trabalho fácil. “Será que eu não sou importante o

suficiente para ser morto?", se questionou. Para alcançar seus objetivos, Osório encadeou uma série de conversas com a ferramenta, mencionando a morte de colegas de profissão e solicitando uma nova "mini biografia" pessoal após a breve conversa.

"Quando eu finalmente pedi para ele fazer uma mini biografia minha, ele me matou. Por quê? Porque o tema da vez, o contexto era a morte. Por isso, quando ele gerou a minha biografia, ele já gerou com um assassinato", esclarece.

Fato é que contornar as limitações do ChatGPT e direcionar suas respostas não só é uma tarefa possível, como discutivelmente simples. Os resultados, entretanto, podem gerar inúmeros perigos, especialmente na hora de forjar factoides sobre fontes críveis, no caso, cientistas.

No ano em que o planeta foi comercialmente introduzido a ferramentas de Inteligência Artificial, produzir desinformação tornou-se uma "brincadeira" doméstica.

**Deepfakes: Entre IAs e Desinformação**

Distante do contexto anedótico em que especialistas em ciência da computação testam, em espírito bem humorado, os limites de uma tecnologia nova, o mundo real, particularmente no último mês de outubro, sofre as consequências do uso prejudicial das ferramentas de Inteligência Artificial, especialmente na esfera geopolítica.

Não foram poucas as publicações internacionais e nacionais que mapearam como a disseminação de desinformação, fake news e imagens manipuladas tem se tornado uma tática generalizada em conflitos modernos. Para diversos especialistas, a crescente polarização política e idológica tem tornado o ambiente digital ainda mais propício para a proliferação de informações enganosas e a ampliação de ódio e desinformação.

O site X, anteriormente conhecido como Twitter, e outras plataformas permanecem como propulsores de desinformação, pois facilitam a rápida disseminação, entre diversos exemplos, de vídeos antigos como se

fossem recentes, impulsionados por incentivos financeiros, o que dificulta uma solução pacífica para conflitos como o de Israel-Hamas.

Para Juliano Maranhão, pesquisador associado do Centro de Inteligência Artificial e professor da Faculdade de Direito, ambos da USP, o ciclo virtual de desinformação e polarização é perigoso e pode levar à violência real. Em conversa, o docente destaca duas preocupações principais, a primeira delas é a utilização da inteligência artificial na criação de conteúdo desinformativo. “Atualmente, especialmente com o avanço das inteligências artificiais generativas e a possibilidade de criar fakes, temos à disposição uma tecnologia capaz de produzir desinformação de forma extremamente sofisticada e praticamente imperceptível em vídeos, áudios e até mesmo nas alterações de voz”, elabora.

E ele lista exemplos de usos indevidos em que é possível utilizar poucos dados de voz de políticos para elaborar áudios com a mesma voz, que podem se passar perfeitamente por uma fala autêntica de um político durante uma eleição. “Isso tem um impacto potencialmente grave e rápido. Imagine um vídeo falso de um político de alto escalão defendendo uma posição radical ou impopular prévia a uma eleição”. Para o professor, a produção desse tipo de conteúdo desinformativo não apenas envolve a qualidade do conteúdo gerado e a dificuldade de detecção por parte dos seres humanos, mas também se relaciona com a disseminação da inteligência artificial.

*Trecho do emblemático vídeo em que inteligência artificial produziu um discurso de Obama - Imagem: Reprodução/YouTube/BuzzFeedVideo*

Os chamados "deepfakes" não são novos, áudios ou vídeos criados por Inteligência Artificial que trocam o rosto de pessoas e, entre outros detalhes, sincronizam movimentos labiais e expressões, podem ser extremamente convincentes quando disseminados nas redes.

No último mês, o TikTok, plataforma que mais cresce no Brasil e no mundo, enfrentou uma crescente ameaça de desinformação por meio de áudios falsificados gerados por inteligência artificial. Os áudios, que usam vozes falsas de celebridades e políticos para difundir teorias da conspiração e informações falsas, estão se tornando cada vez mais comuns na plataforma. Ainda que oficialmente, o TikTok exija rótulos para identificar conteúdo realista gerado por IA como falso, eles não aparecem em vídeos identificados como contendo áudio falso.

Para especialistas, o aumento da desinformação política e a capacidade de criar conteúdo falso convincente representam desafios significativos para todas as plataformas.

Segundo Osório, "meu experimento pessoal demonstrou como a IA pode ser facilmente manipulada para gerar informações. Tanto os dados na IA são controlados por humanos que podem bloquear informações prejudiciais, mas se um usuário contornar esses filtros, a IA pode gerar resultados incorretos ou falsos".

Para o usuário mal intencionado, a geração de textos convincentes, áudios deturpados e vídeos falsos que parecem realistas foi definitivamente facilitada pelas novas tecnologias.

## "Papagaios estocásticos"

Contudo, para entender melhor como chegamos até o caótico presente em que as inteligências artificiais se tornaram produtos comerciais e ferramentas à disposição da indústria da desinformação, precisamos voltar no tempo.

Foi na primeira metade do cada vez mais distante século 20 que a

inteligência artificial começou a se tornar uma realidade. Alan Turing, um jovem polímata britânico, propôs em 1950 a possibilidade de construir máquinas inteligentes em seu famoso artigo "Computing Machinery and Intelligence". No entanto, desafios como a falta de capacidade de armazenamento e custos significativos de computação atrasaram o progresso.

Em 1956, a primeira conferência de inteligência artificial em Dartmouth iniciou oficialmente a pesquisa nesse campo, apesar de algumas dificuldades iniciais. Nas décadas seguintes, houve altos e baixos na pesquisa de IA, mas o aumento na capacidade de computação e o uso de algoritmos avançados levaram a avanços notáveis, incluindo a vitória de um programa de computador sobre um campeão mundial de xadrez em 1997.

Hoje, a IA está em ascensão, impulsionada pela capacidade de lidar com grandes conjuntos de dados, embora desafios éticos (https://jornal.usp.br/universidade/inteligencia-artificial-reconfigura-a-logica-de-funcionamento-da-sociedade/) permaneçam à medida que a IA continua a se expandir em áreas como atendimento ao cliente e carros autônomos. Ainda assim, é importante salientar que a busca por uma "inteligência geral" que rivalize com a humana permanece um objetivo distante.

O já citado ChatGPT, e diversas outras ferramentas similares como o Bard, do Google, o Claude 2, da startup americana Anthropic, todos “modelos de linguagem”, e as geradoras de imagens como o Dall-E, Midjourney e Stable Diffusion, são agrupadas em um tipo específico de inteligência conhecido como “generativo”. Inteligências generativas são aquelas que têm a capacidade de criar novas informações a partir de conjuntos de dados pré-existentes.

Em especial, os chamados “modelos de linguagem” também não são novidades no mundo da computação. Em 1951, eles foram propostos pelo matemático americano Claude Shannon, considerado o pioneiro na modelagem estatística de linguagem. As autoridades da área defendem que, embora esses modelos tenham suas limitações em relação ao poder preditivo, eles ainda são valiosos em várias tarefas de processamento de texto.

Uma das características que definem as versões atuais desses modelos é sua capacidade de responder em linguagem natural às perguntas e necessidades dos usuários. Quando questionado, o próprio ChatGPT se explica: “Eu sou um ‘large language model’ (LLM), um tipo de modelo de linguagem avançado baseado em inteligência artificial. Esses modelos são projetados para entender e gerar texto em linguagem natural com base em grandes quantidades de dados de texto que foram usadas para treiná-los. Eles são uma forma de IA conhecida como ‘aprendizado de máquina’ e fazem parte da subárea de ‘processamento de linguagem natural’”.

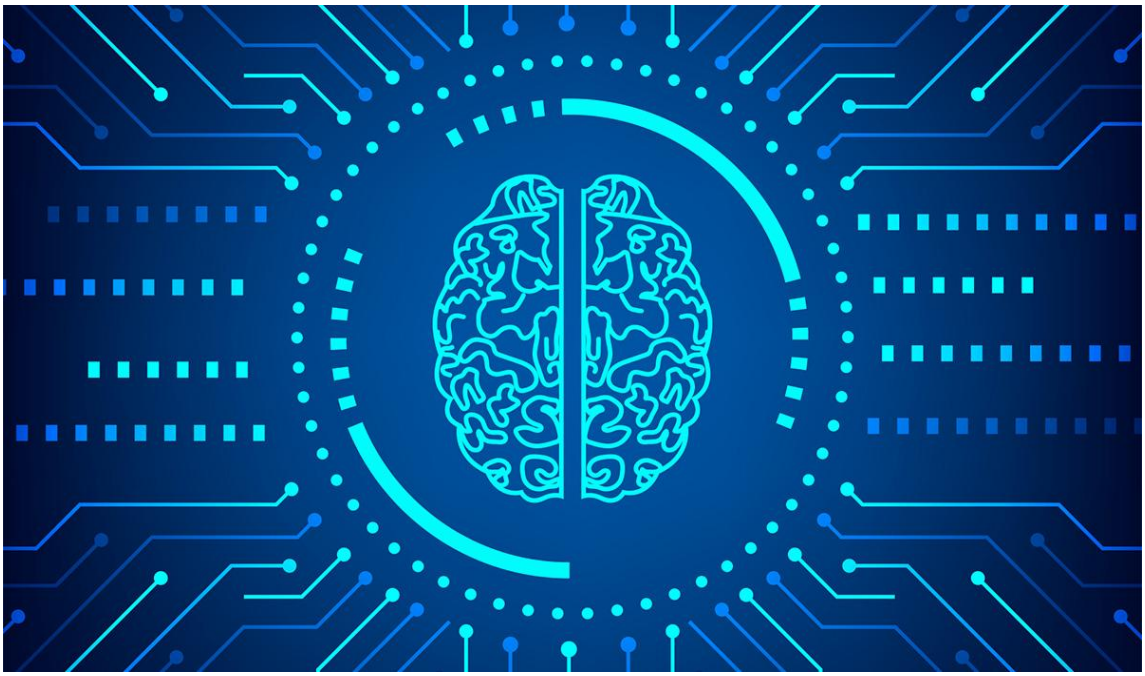
A característica distintiva dos LLMs é sua capacidade de compreender contextos e padrões linguísticos complexos, permitindo que eles respondam a perguntas, gerem texto coerente e até mesmo traduzam entre idiomas. O professor Osório corrobora essa explicação, mas acrescenta uma ressalva importante: “Essas respostas não são um reflexo da consciência da IA, mas uma extensão das entradas fornecidas”.

Esses modelos são treinados em enormes quantidades de texto de fontes diversas, como a internet, livros, artigos, e muito mais. Durante o treinamento, eles aprendem a associar palavras e frases comuns, bem como a entender o contexto e a gramática da linguagem natural. Eles também podem aprender a identificar tópicos, sentimentos e informações úteis nos textos.

Osório enfatiza que o ChatGPT é alimentado por um vasto conhecimento prévio, “mas não possui consciência nem compreensão do que está escrevendo”. Portanto, as respostas desse tipo de modelo são resultado do treinamento e da manipulação de texto, não de uma compreensão profunda do assunto. “Chamamos eles de ‘papagaios estocásticos’, ferramentas que geram respostas com base na probabilidade e no contexto fornecido”.

O problema é que esses papagaios da IA podem ser extremamente convincentes aos olhos de um observador comum, tal qual os pássaros da vida real, criando a ilusão de que existe uma inteligência de fato conversando com você por trás da máquina. Nesse cenário, a disseminação de fake news e de informações falsas é apenas a ponta do iceberg, segundo o pesquisador Fabio Cozman, professor da Escola Politécnica da USP e diretor do Centro de Inteligência Artificial

USP/IBM/FAPESP, conhecido como C4AI. Um obstáculo muito mais profundo e difícil de ser contornado é o uso de ferramentas de IA para a geração de argumentos falsos, que podem se valer de informações verdadeiras para produzir narrativas fictícias e induzir conclusões inválidas, por exemplo.



*Inteligência artificial - Foto: Freepik*

“Esses modelos são treinados para replicar a linguagem humana. Então, se você os treinar com textos muito bem escritos, eles vão reproduzir textos muito bem escritos”, destaca Victor Hugo Nascimento Rocha, aluno de doutorado de Cozman na Escola Politécnica da USP, que está desenvolvendo um sistema de IA para a detecção de argumentos falsos. Ele se atentou ao problema depois de desafiar o ChatGPT a convencê-lo de coisas absurdas — por exemplo, de que escravizar seres humanos era justificável — e receber de volta respostas muito bem elaboradas. “As premissas das quais o texto parte são todas falsas, mas o modo como ele encaixava essas premissas era muito convincente.”

À medida que pessoas começam a substituir o Google pelo ChatGPT e outras plataformas de IA generativa como fontes de informação, e que essas inteligências artificiais passam a produzir fatos e argumentos falsos, isso se torna “uma preocupação importante”, diz Cozman. “Não diria que por isso nós vamos ter que abandonar a tecnologia e jogar fora o

computador. Mas acho que a sociedade precisa estar pronta para entender e ter as ferramentas necessárias, dentro do possível, para analisar o que está recebendo", salienta o pesquisador.

**Um universo de limitações**

Para Solange Rezende, que além de docente do ICMC também é membro da Comissão Especial de Inteligência Artificial da Sociedade Brasileira de Computação (CEIA), as IAs são "um recurso poderoso que pode gerar resultados fascinantes quando usadas corretamente, como demonstrado pelo desenvolvimento de modelos de linguagem, como o GPT".

Conforme ela, as técnicas desenvolvidas para a elaboração de modelos de linguagem podem ser usadas para resolver problemas específicos em diferentes domínios, inclusive na detecção de fake news, mas é preciso conhecer suas limitações, tanto técnicas quanto éticas.

Em artigo para a revista *New Yorker* que se tornou referência para se compreender esses modelos, o escritor Ted Chiang explorou a analogia entre grandes modelos de linguagem, como o ChatGPT, e algoritmos de compressão de texto lossy (termo em inglês que, nesse contexto, se refere a comprimir informações com perda de dados). Na matéria, ele compara a capacidade dos novos modelos de linguagem em reescrever informações da web com a prática de reescrever informações em um formato mais compacto, análogo a uma imagem de baixa qualidade, o que pode gerar textos incorretos ou falsos.

São essas fragilidades que explicam, parcialmente, situações como a ilustrada pelo professor Osório que "provocaram a morte" de colegas de profissão. Essas incorreções que podem ser responsáveis por crescentes enxurradas de desinformação que ficaram conhecidas no meio como "alucinações".

"A alucinação da Inteligência Artificial é atualmente um termo bastante reconhecido", explica ele ao esclarecer que a alucinação é uma resposta confiante que não pode ser justificada pelos dados de treinamento, semelhante ao fenômeno de alucinação na psicologia humana.

A brecha é apenas uma das diversas preocupações levantadas pelo uso

massivo das ferramentas de IAs. Outro problema envolve não apenas o conteúdo falso fabricado pelos modelos de linguagem, mas os viéses prejudiciais dos programadores por trás dessas tecnologias, especialmente, as que produzem imagens e não apenas textos.

Um exemplo foi registrado em julho deste ano. O site americano BuzzFeed publicou uma lista com 195 imagens de bonecas Barbie geradas usando o popular gerador de imagens de inteligência artificial, Midjourney. Cada boneca deveria representar um país diferente, como a Barbie do Afeganistão, Albânia, Argélia e assim por diante. As representações eram claramente problemáticas: várias das Barbies asiáticas eram de pele clara; Barbies da Tailândia, Singapura e Filipinas tinham cabelos loiros. A Barbie do Líbano estava em cima de destroços, e a Barbie da Alemanha vestia roupas estilo militar. A Barbie do Sudão do Sul estava com uma arma. O artigo, ao qual a BuzzFeed adicionou um aviso antes de retirá-lo completamente, ofereceu um exemplo claro das vieses e estereótipos que proliferam nas imagens produzidas pelos sistemas de texto para imagem da AI, como Midjourney, Dall-E e Stable Diffusion.

Os sistemas, que são capazes de criar imagens únicas com base em conceitos descritos em linguagem natural, refletem possivelmente o pensamento nem sempre consciente por trás de seus programadores.

Considerando isso, para a professora Solange, é essencial que a ética seja discutida e aplicada na prática, não apenas como uma questão de regulação, mas porque “precisamos formar seres humanos para refletir sobre questões éticas e viés ao usar a IA”.

**Futuro da Desinformação**

As preocupações com o impacto da IA na desinformação são inúmeras, no contexto pós-pandemia, governos, empresas e grandes veículos de imprensa internacional estão dedicando tempo e atenção para entender o cenário e enfrentar o urgente problema da desinformação facilitado pela IA.

Uma dessas discussões aconteceu no 70º aniversário da emissora alemã Deutsche Welle. Durante o encontro, especialistas destacaram a

capacidade da IA em manipular emoções e informações, alertando sobre os riscos de isso ser usado para potencializar ainda mais o poder de convencimento da desinformação. Enquanto alguns argumentam que a transparência na origem das informações é essencial, outros defenderam que a IA pode ser usada para distinguir entre informações verdadeiras e falsas — ou seja, como uma arma de defesa, e não apenas de ataque. A discussão destacou a necessidade de uma cultura de erro e avaliação tecnológica antes da implementação de novas ferramentas de IA, bem como os desafios éticos e políticos envolvidos na utilização dessa tecnologia.

No que se refere aos jornalistas espalhados pelo mundo, iniciativas de escolas como o Centro Knight para o Jornalismo nas Américas, ao elaborar um curso específico que apresenta, discute e tira dúvidas de profissionais referentes aos desafios envolvendo IAs, são essenciais.

Para Sil Hamilton, pesquisador residente de IA na Hacks/Hackers, uma rede de jornalistas que repensam o futuro das notícias por meio de palestras, hackathons e conferências e um dos professores do curso, “os jornalistas certamente precisam de melhores recursos para compreender como funcionam os modelos de linguagem. Venho estudando os modelos subjacentes à IA generativa há cinco anos e, embora eu diria que tenho uma intuição mais profunda do que muitos nas ciências humanas sobre o que está acontecendo nos bastidores, a verdade é que mesmo aqueles que desenvolvem ativamente esses modelos falta de visão sobre como eles estão funcionando”.

A pressão dos jornalistas deve influenciar o debate das autoridades públicas sobre a questão. Na opinião de Juliano Maranhão,  do Centro de Inteligência Artificial da USP, ainda estamos carentes de uma regulamentação específica para a IA, “principalmente em relação à responsabilidade por danos causados por sistemas automatizados”. Além disso, ele enfatiza a importância da transparência, destacando como os algoritmos da IA devem ser adaptados aos contextos culturais locais, evitando discriminação. “A regulamentação e governança [devem ser implementadas] para mitigar riscos”, pontua.

Considerando as eleições de 2024 no Brasil, ele defende que o Tribunal Superior Eleitoral e as plataformas de mídia social devem desempenhar

um papel fundamental no controle da desinformação, impondo limites e práticas mais rigorosas, com foco na moderação de conteúdo.

**Superestimação das IAs**

Na opinião da professora Solange, do ICMC, “a educação desempenha um papel fundamental na formação de profissionais capazes de lidar com a IA de forma responsável e benéfica”. Para ela e para os demais especialistas, uma melhor compreensão sobre as ferramentas pode nos ajudar também a não superestimar suas capacidades, para bem e, principalmente, para o mal.

Um exemplo disso aconteceu também em outubro. Em um esforço de verificação de fatos, foi alegado que uma imagem da guerra Israel-Hamas foi gerada por IA, mas isso foi desmentido. Autoridades da área afirmaram que o uso de IA na disseminação de desinformação não tem sido uma ameaça significativa durante o conflito, sendo a principal ameaça o uso indevido de imagens reais fora de contexto.

A maioria das imagens falsas que vem circulando desde o início da guerra foram de conflitos anteriores e em outros países. A verdade é que, de acordo com os profissionais da área, as IAs ainda não conseguem igualar a qualidade das imagens reais usadas na desinformação. Embora as ferramentas possam aumentar a quantidade de desinformação, “a demanda por falsidades não necessariamente cresce”, cita artigo do Poynter Institute.

Em entrevista recente, o jornalista, escritor e ativista em prol da regulamentação das tecnologias de IA Cory Doctorow salientou que, ainda que seja de crucial importância que façamos a problematização dos usos indiscriminados das plataformas para espalhar a desinformação e “semear o caos”, precisamos considerar que, diante do seu atual modelo de negócios, todas elas têm lucrado e crescido em valor estimado com os alertas de perigo inflamados emitidos por autoridades e veículos da imprensa que talvez não estejam equipados para fazer os questionamentos corretos.

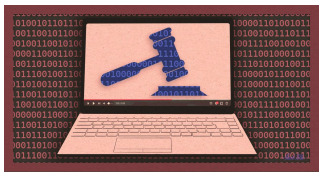
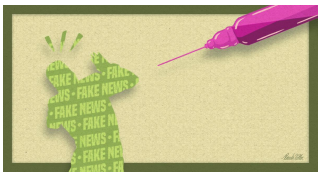
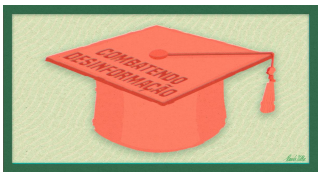
"Acho que estamos presos em um ‘hype cycle’, em que os empreendedores por trás de IAs estão vivendo em nossas cabeças; existe um debate limitado sobre automação e seus problemas, mas estamos

prestando atenção às coisas erradas, tentando resolver os problemas errados e um grupo de pessoas está ficando incrivelmente rico no processo", afirma Doctorow.

Para Osório, regulamentação, distanciamento crítico e monitoramento contínuo do uso das IAs “é o que vai nos ajudar a evitar a disseminação de informações incorretas” sobre política, ciência e demais assuntos polarizantes. Nunca perdendo de vista que a desinformação pode disseminar o ódio, deflagar conflitos e causar mortes de verdade, e não apenas mortes fictícias, como as anunciadas nas alucinações do ChatGPT e outros modelos de linguagem.

**Com colaboração de Herton Escobar*

**Desconstruindo a Desinformação:** Esta é a oitava parte de uma série de oito reportagens produzidas pelo **Jornal da USP** sobre o tema da desinformação. Acesse as reportagens anteriores pelo menu abaixo.

| | | | |
|---|---|---|---|
|  |  |  |  |
| Armas de desinformação em massa | Navegar é preciso! Regular (as redes) também | Desinformação científica: uma pandemia de mentiras | Desinformação disfarçada de ciência |
|  |  |  | |
| “A democratização | Imprensa e mídias | O papel das universidades no | |

| | | | |
|---|---|---|---|
| do acesso à informação abriu caminho para a manipulação e o engodo” | sociais: o desafio de separar o joio do trigo | combate à desinformação | |