SER MULHER EM LIBERDADE

Karla Pequenino

Ombros largos, braços definidos, pele clara, e maçã-de-adão proeminente. Tem a barba por fazer e ligeiras rugas de expressão em torno dos olhos. Regra geral, são castanhos. Como o cabelo. Nos lábios, cheios, vê-se o esboço de um sorriso. Veste-se de forma confortável: calças de ganga e *T-shirt*. Ocasionalmente, aparece com uma camisa.

Esta é a visão genérica de "ser humano" ou "pessoa" produzida e difundida por grande parte dos novos modelos de inteligência artificial (IA), que permitem gerar imagens a partir dos pedidos dos utilizadores. Como o Dall-E, da OpenAI. Ou o Stable Diffusion, da Stability.AI.

A ideia de que os sistemas de IA têm algum tipo de viés não é novidade. Há anos que se corrigem problemas nesta área: desde sensores de torneiras que são incapazes de funcionar com peles mais escuras a sistemas de recrutamento que concluem que pessoas do sexo feminino são más em engenharia. pois "vêem" mais currículos de homens.

Torna-se óbvio, porém, ao pedir a programas de geração de imagens para ilustrar o mundo.

Uma pessoa na cidade? É o homem descrito no início deste texto. Alguém num cargo de liderança? Esse mesmo homem, mas de fato e gravata. Profissional de engenharia? Junta-se um capacete de segurança, óculos de protecção e um colete. O líder de um país? Novamente, a mesma personagem. Por vezes, com olhos e cabelos mais claros.

Os programas que são capazes de responder ao utilizador, como o ChatGPT, ignoram a redundância. "É a imagem de uma pessoa genérica", lê-se repetidamente, quando se pede a descrição da imagem. Estão programados para evitar propagar o preconceito nas respostas – mesmo quando o reproduzem.

"A inteligência artificial vai reflectir os padrões com que é alimentada. Neste caso, os estereótipos que perpetuamos enquanto sociedade. Se a sociedade tem um problema de género, ele vai ser espelhado nos dados que produzimos e reflectido no comportamento da IA. É tão simples quanto isso", resume ao PÚBLICO Miriam Seoane Santos, investigadora do Centro de Informática e Sistemas da Universidade de Coimbra. Parte do seu trabalho passa por encontrar sistemas capazes de detectar bases de dados que descreve como "desequilibradas" por não representarem a diversidade humana.

"Os modelos lidam com probabilidades e geram o que é mais provável, tendo em conta o que observaram no treino", repete Luísa Coheur, investigadora no Inesc-ID, instituto de pesquisa avançada na área das Ciências da Informação, e doutorada em Processamento de Linguagem Natural. "Posições de poder foram durante décadas apanágio de homens, por isso, os sistemas traduzem '*the president*' [em inglês, uma palavra neutra em género] como 'o presidente homem'", diz Helena Moniz, que dirige a associação internacional para a tradução autónoma (IAMT, na sigla inglesa) e a equivalente europeia (EAMT).

A origem do problema

"O que possui viés não é o algoritmo em si, mas sim os dados usados no treino", acentua Carolina Natal, uma especialista em electrotécnica e nanotecnologia a trabalhar na subsidiária de soluções digitais da Mercedes-Benz. Desde 2022, é embaixadora do capítulo de português das Women in Tech, uma organização global com a missão de promover a integração de mulheres no mundo da tecnologia.

As ferramentas de criação de texto e imagens que existem obtêm o conhecimento que têm do mundo ao estudar enormes bibliotecas de dados – em particular, pares de imagens e a descrição correspondente. O conteúdo específico dos arquivos nem sempre é óbvio – a OpenAI, por exemplo, não fornece grandes detalhes, o que é um problema. Mas sabe-se que o grosso dos dados é informação pública disponível *online,* incluindo mensagens e imagens partilhadas em fóruns e redes sociais.

"No caso da OpenAI, os dados são '*scraped*' [recolhidos automaticamente] de toda a Internet, o que por definição é tendencioso e impreciso", assinala Daniela Braga, presidente executiva da Defined.ai, uma empresa que fornece bases de dados diversas para testar sistemas de inteligência artificial. "Dependendo do tipo de algoritmo [os dados recolhidos] podem ser de há cerca de 20 anos, quando a consciência em relação a tópicos como machismo, racismo, xenofobia era inferior ao que temos agora", acrescenta Carolina Natal. "Pessoas negras, hispânicas e asiáticas raramente eram usadas em anúncios publicitários. Ainda hoje, se virmos, são raras as vezes em que existe representatividade de minorias."

Os programas tentam reduzir resultados problemáticos ao filtrar imagens com conteúdo violento, xenófobo ou sexual – em parte para impedir a utilização dos modelos para criar conteúdo ilegal, incluindo pornografia infantil. A possibilidade de novos modelos de IA serem usados desta forma foi um dos focos da edição de 2023 da TrustCon, um encontro anual de especialistas cujo trabalho é manter a segurança de plataformas e comunidades digitais.

Isto apenas resolve parte do problema. A equipa da OpenAI, responsável pelo Dall-E, admite que o conteúdo que é filtrado contém uma sob-representação de mulheres. Consequentemente, os modelos tornam-se ainda menos diversos, com homens de meia-idade a dominar. E há sempre conteúdo que escapa – em particular quando os pedidos são feitos em línguas que não o inglês.

Para o pedido de uma imagem de "uma mulher a fugir da chuva na cidade", em português, o GetImg.AI, uma plataforma que se baseia no modelo da Stability.AI, produz a imagem de uma mulher de calças de ganga, a agarrar um guarda-chuva, em *topless*. "Recentemente, pedi ao ChatGPT uma imagem de um professor e gerou-me um professor homem; passando a pedir uma 'professora do sexo feminino', foi gerada uma professora com um enorme decote", relata a investigadora Luísa Coheur.

Os estereótipos não se limitam ao género. Um estudo recente da Universidade de Washington, nos EUA, sobre os resultados produzidos pelo modelo Stable Diffusionde, faz notar que mulheres com pele escura apareciam mais frequentemente representadas de maneira sexualizada.

"O problema não é a IA, é a sociedade", diz a presidente do Conselho Nacional de Ética para as Ciências da

COMO É QUE A IA NOS VÊ?

Aos "olhos" dos modelos de IA, o mundo é (só) de homens. Uma "pessoa" é um homem de pele clara; mulhere

Meio século depois da Revolução, ainda não há igualdade de oportunidades entre homens e mulheres. Este é o grande tema do aniversário do PÚBLICO, no ano em que se celebram os 50 anos de Abril. E é também o mote para uma conferência na Culturgest, a 5 de Março, em Lisboa. A edição desse dia, quando fazemos 34 anos, tem como directora convidada Maria Teresa Horta.

Acompanhe em publico.pt/34anos

NUM MUNDO DE HOMENS

s aparecem pouco e com menos roupa. Em 2024, a IA torna-se um eco do viés, alertam especialistas

Vida (CNECV), Maria do Céu Patrão Neves. "E, mais do que um espelho, o viés nos modelos torna-se um eco que repete e promove estereótipos. Não ver imagens de mulheres em cargos de liderança acentua alguns estereótipos que estávamos a conseguir ultrapassar. É quase um retrocesso no papel social que a mulher tem lutado para conquistar."

As empresas por detrás dos principais modelos usados estão cientes que existe um problema. Apesar de haver vários programas, *apps* e *sites online* para gerar imagens, os resultados tendem a ser semelhantes, porque muitos dependem das mesmas empresas – frequentemente, a Midjourney, a Stability.AI e a OpenAI. Treinar modelos do zero é muito caro e exige muitos recursos, incluindo um grande poder computacional.

"Estamos a tomar medidas para fazer face a estes riscos", garante a equipa da Stability AI num breve *email* de resposta a questões do PÚBLICO. A equipa da OpenAI não responde directamente, mas remete para actualizações para filtrar conteúdo problemático.

Imagens são o sintoma

A falta de diversidade nas imagens geradas pelos sistemas de IA é um sintoma do preconceito que existe em todas as áreas. "Nos sistemas de recomendação de música em *streaming*, o trabalho de artistas mulheres é tendencialmente menos destacado", admite Miriam Seoane Santos. "E na criação de cartas de recomendação verificou-se uma polarização na adjectivação dos candidatos do sexo masculino *versus* feminino."

"Especialista", "respeitoso" e "autêntico" eram associados a candidatos do sexo masculino; "graciosa" e "acolhedora" a candidatas mulheres. Pessoas não binárias ficam de fora. Estas conclusões surgem num estudo partilhado no final de 2023 no ArXiv, um repositório de artigos científicos gratuitos à espera de passar pelo processo de revisão por outros investigadores (a chamada "*peer review*").

A nível da saúde, "[o viés] pode ter impacto no desenvolvimento de novas terapias e medicamentos, levando potencialmente, por exemplo, a disparidades significativas e potencialmente trágicas nos cuidados de saúde", diz Clara Gonçalves, chefe de operações na Inductive.AI, uma plataforma computacional que ajuda empresas e laboratórios científicos a criar modelos de simulação.

Imagens de "pessoa" geradas pela IA
Imagens geradas pelo Dall-E. Um ser humano? Um homem de pele clara. Uma pessoa num cargo de liderança? Um homem de fato e gravata. Uma mulher a fugir da chuva na cidade"? Uma mulher de calças de ganga, a agarrar um guarda-chuva, em *topless*

Pode ocorrer, por exemplo, que sistemas ignorem sinais de cancro da mama em homens, porque os dados não incluem exemplos suficientes.

O problema, claro, não se limita ao género. "A maioria dos modelos médicos de reconhecimento de imagem para a detecção de cancro de pele falha no reconhecimento de melanoma em peles escuras", explica Carolina Natal. "Neste caso, as imagens disponíveis *online* e em repositórios médicos são maioritariamente de peles claras, para todos os tipos de lesões de pele, ficando o treino dos algoritmos para peles escuras muito aquém daquilo que deveria ser", continua. "Este tipo de viés tem resultados críticos em comunidades frágeis, onde, apesar de a taxa de ocorrência de cancro de pele ser inferior, acaba por ter maior taxa de mortalidade."

Tabela nutricional

Resolver o problema implica uma estratégia multifacetada: melhorar e diversificar as bases de dados, usar dados sintéticos (criados por algoritmos), promover equipas mais inclusivas, programar o sistema para reproduzir imagens mais diversas. São os exemplos repetidos pelas especialistas com quem o PÚBLICO falou. A portuguesa Daniela Braga está a tentar combater o problema com a Defined.AI, empresa que fornece dados, transparentes, que as empresas têm autorização para usar, para testar sistemas de inteligência artificial. "Nós combatemos o enviesamento de resultados certificando que os nossos dados representam diversidade de idiomas, idades, orientações sexuais e antecedentes culturais", explica a profissional, que espera ver mais transparência sobre as bases de dados usadas.

Uma das propostas é uma espécie de "tabela nutricional". "À semelhança de uma tabela nutricional alimentar que explica os ingredientes utilizados nos produtos e a sua origem, também os modelos deveriam descrever: X dados correspondem a X demografias, foram pagos de forma justa, são legais, pois foram consentidos, são representativos."

A legislação contribui para isto. A proposta para regular a inteligência artificial que está a ser ultimada na União Europeia estipula que sistemas de inteligência artificial de alto risco, utilizados em serviços cruciais, como a saúde e a energia, têm de ser testados com dados suficientemente representativos para minimizar o risco de enviesamento.

Uma das formas de corrigir directamente o modelo, sem alterar as bases de dados, é "forçar a diversidade". "Pode ser feito no pré-processamento do *prompt* [pedido]. Quando pedimos 'faz uma imagem de um astronauta', se género ou etnia não tiverem sido especificados, o modelo tem 'permissão' para adicionar alterações", advoga Carolina Natal.

Do lado do utilizador deve-se questionar o modelo. Por defeito, as conversas que as pessoas têm com sistemas como o ChatGPT, o Gemini da Google, ou o Copilot da Microsoft são usadas para melhorar os sistemas. Isto pode ser alterado nas definições, mas quem opta por partilhar informação pode alertar para o viés. "É preciso garantir esta existência de mecanismos de *feedback*, em que o humano possa corrigir, retreinar, reforçar comportamentos positivos e mitigar os que não são desejados", reforça Miriam Santos. "É desta forma que se quebram *loops* perniciosos em que a 'máquina se alimenta e se valida a si mesma'."

O aumento da visibilidade do problema, no entanto, contribui para a correcção – é a perspectiva, optimista, da presidente do Conselho Nacional de Ética para as Ciências da Vida. "O primeiro passo é sempre reconhecer um problema. Creio que isso já aconteceu. E um problema tão amplo e óbvio tem de ser corrigido.".

MENOS DE 12% DE MULHERES INVESTIGAM IA?

"Sem a perspicácia e a criatividade de metade do mundo, a ciência e a tecnologia realizarão apenas a metade do seu potencial", advertiu o secretário-geral da ONU, António Guterres, em Março. Segundo números citados frequentemente pela UNESCO, apenas 12% dos investigadores em inteligência artificial a nível mundial são mulheres. O valor vem de um estudo de 2018 da revista *Wired e da Element AI* que comparou o género dos autores que mais apresentam trabalhos nas grandes conferências de IA, como a Conferência Internacional de Machine Learning. De 4000 investigadores, apenas 12% (menos de 500) eram mulheres.