Ética & Inteligência Artificial

Marcelo Fantini RA 108341
Rubens de Castro Pereira RA 217146

25 de março de 2023

1. Defina Ética em Inteligência Artificial.

A ética em Inteligência Artificial (IA) representa um conjunto de valores e princípios que devem ser respeitados por todos os atores envolvidos no ciclo de vida de sistemas de IA com destaque ao respeito, proteção e promoção dos direitos humanos, liberdades fundamentais e dignidade humana, diversidade e inclusão [1].

2. Apresente uma notícia recente de um problema de Ética em IA.

O site de notícias *The Verge* publicou o artigo *Anyone can use this AI generator - that's the risk* [2]. A notícia discute o avanço da inteligência artificial na área de programas de texto-para-imagens. Esses programas abriram a possibilidade para que pessoas sem habilidades artísticas pudessem, com prompts de texto, gerar imagens através da inteligência artificial, que utiliza de uma vasta base de imagens para gerar o pedido.

A inteligência artificial está longe de ser perfeita. Ela possui problemas para gerar mãos, ocasionalmente existem deformidades nas pessoas, entre outras falhas. Entretanto, essas falhas não são incômodas para quem está empolgado com a tecnologia, que pode gerar qualquer imagem que você pode imaginar.

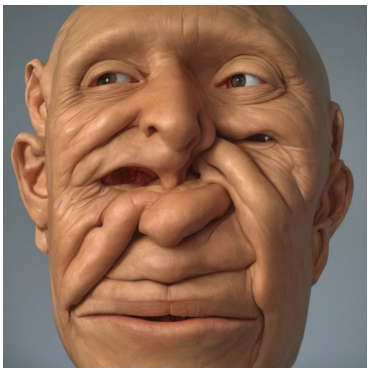

Figura 1: Imagem gerada por um prompt pedindo uma escultura hiperrealista de uma face humana. Encontrada no Lexica.

A empresa OpenAI possui o gerador de imagens *DALL-E*, que possui uma quantidade finita de geração de imagens gratuitas por mês. Após esta quantidade se esgotar é necessário pagar por prompts para gerar novas imagens, criando uma pequena barreira para aqueles que querem gerar imagens. O Google possui um gerador de imagens chamado *Imagen*, mas este não está aberto ao público.

Além desses geradores ganhou notoriedade o método Stable Diffusion, difundido pela empresa Stability AI. A empresa, chefiada pelo CEO Emad Mostaque, foca no desenvolvimento open source de Stable Diffusion. Mostaque diz que a iniciativa open source é sobre "colocar o controle nas mãos das pessoas que construirão e estenderão a tecnologia."Entretanto, isso significa colocar **todas** essas capacidades nas mãos do público, para o bem ou para o mal.

Um dos problemas do Stable Diffusion é que não existem restrições sobre o tipo de conteúdo que pode ser gerador. Outros geradores, como DALL-E e Imagen, possuem restrições severas nas palavras-chave e conteúdo que podem gerar, enquanto que o Stable Diffusion pode ser utilizado localmente.

Uma vez na máquina local de um usuário, não existe como restringir o que é gerado. Isto torna muito mais fácil de gerar conteúdo violento e sexual, incluindo imagens de pessoas reais. Com Stable Diffusion, o caso mais comum até o momento são usuários gerando pornografia.

Essa situação é território essencialmente desconhecido, e não é claro quais serão as consequências de soltar um modelo como esse para o público. É fácil de imaginar os fins maliciosos para os quais essa tecnologia pode ser utilizada, mas isso não significa que essas previsões acontecerão.

Outro problema é o uso de imagens com direitos autorais utilizadas como treinamento e base para as imagens geradas pelo Stable Diffusion. Apesar da empresa Stability AI fazer alguns filtros, ela não filtra o uso de bancos de dados de direitos autorais. Como resultado, muitos vêem a habilidade de Stable Diffusion imitar o style e estética de artistas vivos. Não apenas uma brecha de direito autoral mas também ética.

O aspecto de direitos autorais adiciona uma nova dimensão às reclamações que ferramentas como Stable Diffusion estão tirando trabalhos de artistas humanos. Não apenas está roubando trabalhos de artistas como está fazendo isso através de contrabandear, por assim dizer, as habilidades que esses indivíduos necessitaram de horas e horas para aperfeiçoar.

3. Apresente um artigo científico recente de uma solução para um problema de Ética em IA.

O artigo intitulado *A Pathway Towards Responsible AI Generated Content* [3] tem como propósito tratar dos riscos potenciais e uso indevido de geração de conteúdo por IA (GCIA), auxiliar na eliminação de obstáculos e promover entregas seguras e éticas de conteúdos gerados por IA. Os tipos de conteúdos podem ser bastante variados podendo ser imagens, textos, áudios e vídeos.

Com o uso extensivo da GCIA tem surgido preocupações relativas à privacidade, preconceito, toxicidade, desinformação, propriedade intelectual e potencial uso indevido diante das pessoas. Uma questão relevante surgiu recentemente com a disponibilização de novas funcionalidades no ChatGPT as quais permitem fazer a depuração de código fonte de programas de computador ou elaborar trabalhos escolares/acadêmicos. Esses resultados podem gerar riscos potenciais, pois os modelos de GCIA produzem trabalhos replicando conteúdos com os quais foram treinados devido á elevada capacidade de memorização. O conjunto de dados utilizados para treinamento frequentemente tem origem e direitos autorais desconhecidos, muitas das vezes não passam por uma análise de curadoria cuidados. A maioria dos modelos de GCIA decodificadores de textos são treinados com grandes quantidades de dados obtidos da internet, os quais podem conter desvios (bias) relacionados a temas sociais, podem ser tóxicos, e outras limitações inerentes aos grandes modelos de linguagens.

Para que os modelos de GCIA sejam responsáveis, estes devem considerar o seguinte escopo: privacidade, viés (tendências), toxicidade, desinformação, proteção da propriedade intelectual. Além disso,deve contemplar também a robustez dos sistemas, explicabilidade (feedback), código fonte aberto, consentimento, créditos e compensação, e ambiente amigável para o seu uso. O artigo [3] tem como propósito tratar dos riscos potenciais e uso indevido de geração de conteúdo por IA (GCIA), auxiliar na eliminação de obstáculos e promover entregas seguras e éticas de conteúdos gerados por IA. Os tipos de conteúdos podem ser bastante variados podendo ser imagens, textos, áudios e vídeos.

Com o uso extensivo da GCIA tem surgido preocupações relativas à privacidade, preconceito, toxicidade, desinformação, propriedade intelectual e potencial uso indevido diante das pessoas. Uma questão relevante surgiu recentemente com a disponibilização de novas funcionalidades no ChatGPT as quais permitem fazer a depuração de código fonte de programas de computador ou elaborar trabalhos escolares/acadêmicos. Esses resultados podem gerar riscos potenciais, pois os modelos de GCIA produzem trabalhos replicando conteúdos com os quais foram treinados devido á elevada capacidade de memorização. O conjunto de dados utilizados para treinamento frequentemente tem origem e direitos autorais desconhecidos, muitas das vezes não passam por uma análise de curadoria cuidados. A maioria dos modelos de GCIA decodificadores de textos são treinados com grandes quantidades de dados obtidos da internet, os quais podem conter desvios (bias) relacionados a temas sociais, podem ser tóxicos, e outras limitações inerentes aos grandes modelos de linguagens.

Para que os modelos de GCIA sejam responsáveis, estes devem considerar o seguinte escopo: privacidade, viés (tendências), toxicidade, desinformação, proteção da propriedade intelectual. Além disso, deve contemplar também a robustez dos sistemas, explicabilidade (feedback), código fonte aberto, consentimento, créditos e compensação, e ambiente amigável para o seu uso.

Privacidade Os modelos base e os modelos generativos de conteúdo possuem a vulnerabilidade de ataques de privacidade, a qual pode ser decorrente de dados duplicados nos datasets de treinamento.

Esse comportamento de replicação tem sido extensivamente estudado nesses modelos, podendo levar a resultados de imagens como a combinação de fundo e de objetos de imagens reais do conjunto de imagens de treinamento. Esses resultados levantam questões sobre memorização de dados, propriedade das imagens difundidas, imagens de pessoas reais, reproduzindo dados do treinamento e não novas imagens criadas pelos modelos.

As questões de privacidade ainda não possuem solução definitiva, mas ações vem sendo tomadas para minimizar tais questões. As companhias tem disponibilizado website para fonecer identificação de imagens já treinadas/memorizadas. Outra ação para evitar a duplicação de dados é o uso de técnicas de deduplicação removendo de forma ampla dados duplicados utilizados em treinamento. Além disso, companhias tem adotado medidas para prevenir o compartilhamento de dados sensíveis pelos seus empregados, evitando o uso desses dados em futuras versões de modelos de GCIA. Atualmente as medidas para evitar o vazamento de dados privados são insuficientes e ainda faz-se necessário explorar sistemas confiáveis para a detecção de dados duplicados em modelos generativos e maior investigação na memorização e generalização em sistemas de aprendizado profundo.

Viés/Tendências/Bias - Toxicidade - Desinformação

Os dados de treinamento utilizados nos modelos de Inteligencia Artifical (IA) são obtidos do mundo real, os quais, sem a intenção, podem reforçar estereótipos indesejáveis, excluir ou marginalizar certos grupos de indivíduos, conter dados tóxicos, levando à incitação ao ódio ou violencia a a ofender indivíduos.

O uso de filtros nos dados é uma possibilidade de minimizar tais problemas, contudo eles podem introduzir vieses nos dados de treinamento que podem se propagar nos modelos. A fim de minimizar esses vieses, técnicas de pré-treinamento são utilizadas nos dados antes do treinamento. Outra estratégia adotada é o contíuo treinamento dos modelos de GCIA com as informações mais recentes evitando o surgimento de lacunas de informação e garantindo que os modelos permanecem atualizados, relevantes e benéficos para a sociedade. Ainda existem lacunas para investigação mais profunda desses de vieses, toxicidade e desinformação em todo o ciclo de vida de desenvolvimento dos modelos, apesar de ser uma tarefa desafiadora.

Proteção da propriedade intelectual

Discussão

Conclusão

Referências

[1] Scientific United Nations Educational and Cultural Organization (UNESCO). Recommendation on the ethics of artificial intelligence, 2022.

[2] James Vincent. Anyone can use this ai art generator - that's the risk.

[3] Chen Chen, Jie Fu, and Lingjuan Lyu. A pathway towards responsible ai generated content, 2023.