IA Generativa:
oportunidades, riscos e governança

Virgilio Almeida

18 de maio de 2023

UFMG

ie] Institute of Advanced Studies of the University of São Paulo

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

IA generativa

- A inteligência artificial generativa é um sistema capaz de criar texto, imagens ou outros conteúdos (música, vídeo, voz, etc.) a partir das instruções de um usuário humano. Esses sistemas podem produzir novo conteúdo a partir dos dados de treinamento. Sua performance agora está próxima de produções feitas por pessoas, devido à grande quantidade de dados que tem sido utilizada em seu treinamento.
- Os sistemas usados no processamento de linguagens naturais são conhecidos como modelos de linguagem. A definição clássica de um modelo de linguagem é uma distribuição de probabilidades sobre sequências de símbolos ou palavras. Para isso, esses modelos são treinados em grandes massas de dados textuais.
- ChatGPT é um grande modelo de linguagem (Large Language Model – LLM), com 175 bilhões de parâmetros, treinado em uma imensa base de dados, capaz de aprender de forma autônoma e produzir textos sofisticados, aparentemente inteligentes. Os parâmetros são chave para os algoritmos de aprendizado de máquina, pois representam aquilo que é aprendido pelo modelo com os dados de treinamento.
- Na historia infantil, o rato comeu o: -------

  - Cachorro
  - Gato
  - Queijo
  - Sapato
  - Cinzeiro

Oportunidades

- Aumento de produtividade
  - Profissionais de ``informação'' e conhecimento
  - Jornalismo e mídia
  - Pesquisa científica

- Oportunidades profissionais para alunos
  - Estudo do MIT aponta que acesso à IA generativa aumentou a produtividade dos agentes em 14%, com o maior impacto nos trabalhadores novatos ou menos experientes.

- Oportunidades para professores
  - Planejamento de aula,
  - Apoio a pesquisa
  - Educação/treinamento professional em áreas específicas

- Oportunidade de capacitar os estudantes no convívio de humano-robots.

Oportunidades de pesquisa

- *Warning: These systems have no ethical understanding of the world, have no sense of truth and they are not reliable (Gary Marcus, NYU).*

- Uma das questões mais imediatas para a comunidade de pesquisa é a falta de transparência.
- Desenvolver métodos para melhorar a qualidade dos resultados em função do tamanho dos conjuntos de dados e métodos para calibrar esses modelos com feedback humano.
- Desenvolver regras de responsabilidade sobre resultados.
- Investir em LLMs verdadeiramente abertos;
- Explorar novas aplicações e benefícios e de LLMs e IA.
- Ampliar o debate multidisciplinar.
- Estudar como o viés nos processos e dados usados para treinar as ferramentas de IA generativa pode ter impacto negativo nos resultados.

Riscos éticos e sociais

- Veracidade dos resultados do ChatGPT é um grande desafio de pesquisa
- Impactos sobre o mercado de trabalho, uma vez que mais funções e ocupações podem ser vistas como obsoletas, demandando transformações profundas nesse universo.
- Preocupação ética sobre autoria e plágio.
- Rrisco de enfraquecimento de instituições tradicionalmente vinculadas à produção de informações e conteúdos.
- Possibilidade de enrijecimento de uma fonte privada de informação, que é potencialmente enviesada e raramente submetida a questionamentos políticos.
- Criação de novos oráculos técnicos – associados a poucas fontes de poder fora do país – data colonialismo.
- Impacto na esfera pública, essencial para democracia, como por exemplo mudanças na consulta pública, onde as sugestões seriam gerada por chatgpt.

New York State Office of the Attorney General Letitia James

1. The Broadband Industry's Campaign to Repeal Net Neutrality Rules in 2017 Resulted in Over 8.5 Million Fake Comments to the FCC — Nearly 40% of the FCC's Total — and Over Half a Million Fake Letters to Congress

Rampant Fraud and Limited Oversight Led to the Submission of More Than 8.5 Million Fake Comments to the FCC and More Than Half a Million Fake Letters to Congress

Fake Comments:

How U.S. Companies & Partisans Hack Democracy to Undermine Your Voice

2. The FCC Received Over 9.3 Million Fake Comments Supporting Net Neutrality That Used Fictitious Identities, Most of Which Were Submitted by a 19-Year Old College Student Using Automated Software

Riscos éticos e sociais

- Desinformação em escala: teorias democráticas fizeram um movimento no sentido de reconhecer que não bastava poder se expressar. Era fundamental ser ouvido e considerado: risco de inundação de conteúdo articulado disputando atenção.
- Tecnologias de IA podem tornar mais complexa a responsabilização por falas, o que dificulta a regulação democrática da expressão.
    - O que acontece, por exemplo, se tais dispositivos começam a reproduzir discursos de ódio?
    - Quem deve ser responsabilizado por isso?
    - Como lidar com uma esfera pública tomada por sínteses linguageiras que podem não expressar opiniões de grupos específicos?
    - E se tais sínteses se tornarem dominantes e convencerem as pessoas de visões e desejos que não tinham a priori?
    - Contribuirão elas para a desmobilização de grupos que reivindicam mudança?
    - Poderão elas alimentar revoltas? Como elas alterarão as atividades de lobby político? Em suma, que consequências políticas podem ter esses atos de fala, já que discursos são, em si, formas de ação? E, como a aparente despersonalização do dizer pode afetar as relações políticas em uma democracia.

Desinformação: diferentes patamares

- Ferramenta poderosa para disseminar desinformação em larga escala.
- Criar novas narrativas falsas pode ser feito em larga escala e com muito mais frequência - é como ter agentes de IA contribuindo para a desinformação.
- As políticas da OpenAI proíbem o uso de sua tecnologia para promover desonestidade, enganar ou manipular usuários ou tentar influenciar a política; a empresa oferece uma ferramenta gratuita de moderação para lidar com conteúdo que promova ódio, automutilação, violência ou sexo.
- A ferramenta oferece suporte limitado para outros idiomas além do inglês e não identifica material político, spam, fraude ou malware. O ChatGPT adverte os usuários de que “ocasionalmente pode produzir instruções prejudiciais ou conteúdo tendencioso”.
- ChatGPT tem o potencial de lançar dúvidas sobre todo o ambiente de informações, ameaçando nossa capacidade de distinguir fato de ficção.

Políticas públicas: interação humano-máquina

- Políticas para o mercado de trabalho

  - impactos sobre o mercado de trabalho, uma vez que mais funções e ocupações podem ser vistas como obsoletas, demandando transformações profundas nesse universo. – (call centers, telemearketing 1,4 milhao de pessoas, mulheres no Brasil)

  - Estudo* analisa a introdução gradual de um assistente de conversação baseado em IA generativa usando dados de 5.179 agentes (operadores) de suporte ao cliente.

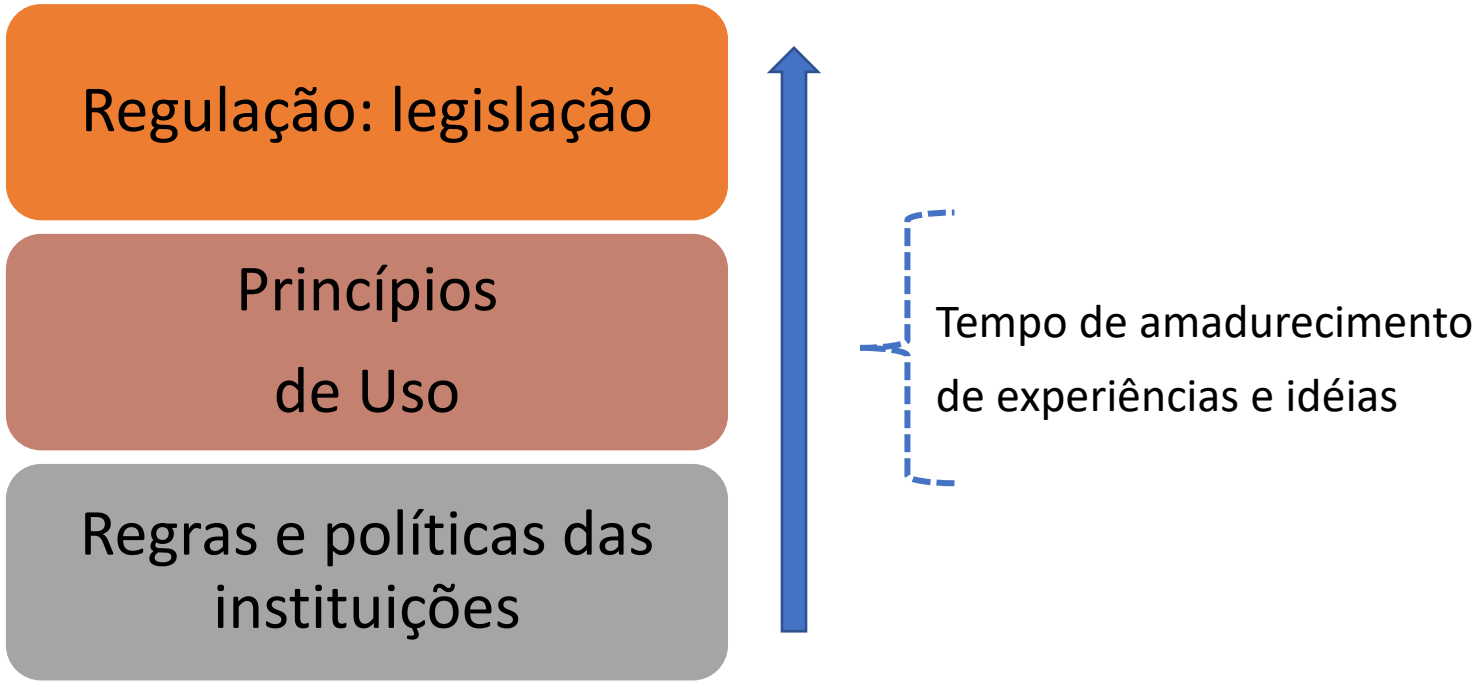
  - O acesso à ferramenta aumenta a produtividade, medida pelo número de problemas resolvidos por hora, em média, em 14%, com o maior impacto em trabalhadores novatos e com baixa habilidade, e impacto mínimo em trabalhadores experientes e altamente qualificados.

  - Fornecemos evidências sugestivas de que o modelo de IA dissemina o conhecimento potencialmente tácito de trabalhadores mais habilidosos e ajuda os trabalhadores mais novos a se adaptarem à curva de experiência.

  - Estudo mostra que assistência de IA melhora a percepeçao do cliente, reduz as solicitações de intervenção gerencial e melhora a retenção de funcionários.

(*) GENERATIVE AI AT WORK, Erik Brynjolfsson et al, Nber Working Paper Series , Working Paper 31161 http://www.nber.org/papers/w31161

Governança: minimizar riscos do IA generativa



Exemplos de Políticas de Uso: ChatGPT e IA generativa

- **Science Magazine**
  - Resumos criados pelo ChatGPT submetidos a revisores acadêmicos; foram detectados apenas 63% dessas falsificações.
  - Fundamento: The work is ``original''.
  - Nova política especifica que texto gerado pelo ChatGPT (ou qualquer outra ferramenta AI) não pode ser usado no trabalho, nem figuras, imagens ou gráficos podem ser produtos de tais ferramentas. E um programa de IA não pode ser um autor.
- **Wired**
  - Não publicamos histórias com texto gerado por IA.
  - Também não publicamos texto editado por IA.
  - *Podemos tentar usar IA para sugerir manchetes ou textos para postagens curtas em redes sociais.*
  - *Podemos tentar usar IA para gerar ideias de histórias.*
  - *Podemos experimentar o uso de IA como uma ferramenta de pesquisa ou análise.*
  - Não publicamos imagens ou vídeos gerados por IA.

Exemplos de Políticas de Uso: ChatGPT e IA generativa

- **PNAS**
  - De acordo com as políticas do PNAS e PNAS Nexus, se um software de IA, como o ChatGPT, tiver sido usado para ajudar a gerar qualquer parte do trabalho, isso deve ser claramente reconhecido; deve ser anotado na seção de Materiais e Métodos (ou Agradecimentos, se nenhuma seção de Materiais e Métodos estiver disponível) no envio.

  - O software não pode ser listado como autor porque não atende aos critérios de autoria e não pode compartilhar a responsabilidade pelo artigo ou ser responsabilizado pela integridade dos dados relatados.

Exemplos de Políticas de Uso em Ensino: ChatGPT e LLM

- **Alunos:**
  - aprender como usar geradores de texto de IA positivamente: aprimoramento pessoal, sem prejudicar, as habilidades de redigir e pensar, honrar preceitos éticos e garantir que avaliações justas, entre quem usa e quem não usa.
  - Estudantes, devem dar crédito aos LLMs sempre que forem usados, mesmo que apenas para gerar ideias em vez de texto utilizável.
  - Incluir um apêndice reportando todas interacoes com LLM no contexto do trabalho. Destacar as seções mais relevantes, escrever uma narrativa no início do apêndice explicando precisamente como foi usado o LLM (para gerar ideias, frases, elementos de texto, longos trechos de texto, linhas de argumentação, peças de evidência, etc.).
  - Explicar a razao do uso do LLM, como economizar tempo, superar o bloqueio da escrita, estimular o pensamento, lidar com o estresse, melhorar a prosa, experimentar por diversão etc.
- **Professor:**
  - deve entender como os LLMs funcionam, incluindo seus pontos fortes e fracos, bem como ferramentas para detectar a saída gerada pelo LLM, a fim de otimizar o valor do aprendizado e incorporar essa entendimento aos procedimentos de avaliação.

Princípios e regulações para uso de IA generativa

- Estados Unidos: em discussão no Congresso.

- Reino Unido e Índia
  - Esses países reconhecem que novas tecnologias não necessariamente exigem novas leis e que os danos de novas regulamentações podem superar quaisquer benefícios potenciais.
  - Reino Unido propõe princípios-chave, como transparência, responsabilidade e reparação.
  - Índia propões mais pesquisas para abordar questões relacionadas à transparência, privacidade e viés e o aprofundamento de pesqusias sobre ética em IA.
- China
  - A Administração do Ciberespaço da China (CAC) propõe a proteção do consumidor e a segurança nacional como motivação para as novas regras de mídia sintética.
  - Aumentar a confiança na IA - que já está entre os níveis mais altos globalmente - o que impulsiona a aceitação pelos consumidores e estimula o crescimento.

Exemplos de Regulação de Uso: China

- Aumentar a confiança na IA - que já está entre os níveis mais altos globalmente - o que impulsiona a aceitação pelos consumidores e estimula o crescimento.
- Os novos regulamentos do governo chinês entraram em vigor em jan. 2023. A Administração do Ciberespaço da China (CAC) cita a proteção do consumidor e a segurança nacional como motivação para as novas regras de mídia sintética, que poderiam inclusive definir um modelo usado em outros países.
- Isso inclui texto, imagens e vídeos gerados por IA. Os dois elementos centrais das regras são verificação e gerenciamento de conteúdo. As plataformas de mídia sintética precisam incorporar algum tipo de sinal de que são produzidas por IA, como uma versão indelével da barra de cores na parte inferior das imagens criadas por DALL-E ou de outra forma “marcadas com destaque para evitar confusão pública ou identificação incorreta”.
- "A regulamentação da IA é altamente iterativa. Há pouca chance de se acertar completamente a regulamentação da IA na primeira tentativa”. (*)
- As autoridades vêm implementando regras a cada poucos meses desde 2021, incluindo uma lei nacional de privacidade, a Lei de Proteção de Informações Pessoais, e um Código de Ética para a Inteligência Artificial de Nova Geração, que estabelece requisitos.
- Ao agir rapidamente na regulamentação, Pequim está criando uma base para exportações de IA para países do Sul Global.

* Source: Decoding China’s Ambitious Generative AI Regulations, Sihao Huang and Justin Curl, Freedom to Tinker, April 2023.

Planos para regulação de IA Generativa na China

- Artigo 4: A prestação de produtos ou serviços de inteligência artificial generativa deve cumprir os requisitos das leis e regulamentos, respeitar a moral social, a ordem pública e os bons costumes, e atender aos seguintes requisitos:
    - Os provedores de IA generativa devem tomar medidas ativas para prevenir a discriminação por raça, etnia, fé, gênero e outras categorias.
    - Os provedores de serviços de IA não podem usar seus algoritmos, dados ou plataforma para se envolver em concorrência desleal.
    - O conteúdo gerado pela IA deve ser "preciso e verdadeiro", e medidas devem ser tomadas para evitar a geração de informações falsas.
    - A IA generativa não deve prejudicar a saúde mental das pessoas, infringir direitos de propriedade intelectual ou infringir o direito à publicidade (ou seja, a imagem de alguém).
- Artigo 5: Indivíduos e organizações que usam modelos de IA generativa para fornecer serviços serão responsabilizados legalmente por conteúdos que violem essas regulamentações.
- Artigo 7: Requisitos rigorosos para dados de pré-treinamento.
- Artigo 15: Para gerações não conformes, além de tomar medidas como filtros de conteúdo, o treinamento de otimização do modelo deve ser usado para evitar a regeneração dos mesmos problemas dentro de três meses.

Desafios Futuros da IA Generativa

- A nova ciência e tecnologia são dinâmicas, de uso duplo e ultrapassarão a regulamentação. É um recurso, não um bug da inovação. Os formuladores de políticas não devem se esforçar para manter a tecnologia constante. O que pode ser mantido constante é um firme compromisso com os direitos fundamentais. **(Alondra Nelson)**

- O futuro reserva os desafios de compreender o alcance das consequências dessas ferramentas no cotidiano dos cidadãos e as implicações para o funcionamento das democracias mundo afora.

- Resta a governos, parlamentos e sistemas judiciais criarem parâmetros e limites, não com base apenas em aspectos éticos, mas, principalmente, sobre as consequências sociais e políticas de tais tecnologias. Para que as tecnologias digitais sejam usadas com responsabilidade é necessário estabelecer políticas e regulamentações para proteção da sociedade contra consequências sociais adversas (Ciência Hoje, Mar/2023)

Obrigado!

Virgilio@dcc.ufmg.br

vafalmeida@gmail.com

Twitter: @virgilioalmeida
Linkedin: virgilio-almeida-872bbb6