SCIAS
Educação, Comunicação e Tecnologia
Education, Comunication and Technology

Três faces do ChatGPT: imaginários de uma máquina de linguagem

Juliana Michelli da Silva Oliveira[1]
Rodrigo de Almeida Siqueira[2]
Rogério de Almeida[3]

Resumo

Com o objetivo de contribuir ao campo de estudos dos imaginários dos objetos técnicos, o presente artigo buscou mapear e problematizar as imagens condensadas no ChatGPT (*Generative Pre-trained Transformer*), modelo de linguagem da *start-up* OpenAI. Para isso, ancorado na perspectiva teórica da antropologia do imaginário e tendo por metodologia a hermenêutica simbólica, o estudo das imagens veiculadas pelo ChatGPT envolveu os seguintes procedimentos: o exame do contexto de produção do artefato; a análise do texto de lançamento do sistema e dos princípios que regem seu funcionamento; a realização de diálogos experimentais com o ChatGPT (denominados de DiEx); além de levantamento em literatura especializada. A partir disto, esta pesquisa exploratória identificou e caracterizou três faces que emergem do ChatGPT: científica, alucinatória e modulável pelo usuário. Por fim, os resultados obtidos apontam para a ascensão da criatividade aumentada, embora envolva também riscos quanto ao empobrecimento de narrativas pela retroalimentação de dados de um universo circunscrito. Tal cenário suscita a necessidade de se refletir sobre as implicações das imagens consteladas pelo ChatGPT e a importância da diversificação dos imaginários na era da criatividade aumentada.

Palavras-chave

ChatGPT; Inteligência Artificial; Antropologia do imaginário.

Recebido em: 08/09/2023
Aprovado em: 26/12/2023

[1] Doutora em Educação pela Universidade de São Paulo (USP), com estágio de pesquisa PDSE no Centre de recherche Imaginaire et Socio-Anthropologie da Université Grenoble Alpes (França). Pesquisadora associada ao Centre de recherche sur le texte et l'imaginaire (Figura), Departamento de Estudos Literários da Université du Québec à Montréal (Canadá). Atualmente é professora no Centro de Estudos Latino-Americanos sobre Cultura e Comunicação (Celacc) da Escola de Comunicações e Artes da USP (ECA-USP). Contato: jumioliveira@gmail.com.

[2] Cursou Engenharia Elétrica na Universidade de São Paulo (Poli-USP). Trabalha com tecnologias de inteligência artificial desde 1995. É sócio fundador da Inbot, onde realiza pesquisa e desenvolvimento de aplicações com inteligência artificial, incluindo sistemas de reconhecimento de linguagem natural, busca semântica, criação de chatbots, linguística computacional, visualização de dados e análise estatística de informações. Contato: rodrigo@inbot.com.br.

[3] Professor Titular da USP – Faculdade de Educação. Bolsista Produtividade CNPq. Coordenador do Lab_Arte. E-mail: rogerioa@usp.br.

Imaginaries of ChatGPT: Three Faces of a Language Machine

Abstract

Intending to contribute to the field of study of the imaginary of technical objects, this article sought to map and problematize the images condensed in ChatGPT (Generative Pre-trained Transformer), a language model from OpenAI, a research and development company. For this, we adopted the theoretical perspective of the anthropology of the imaginary and, employing symbolic hermeneutics as the methodology, we studied the images conveyed by ChatGPT following multiple-method procedures, such as examining the context in which the artifact was produced, analyzing the system's launch text, as well as the principles that govern its operation. We also conducted experimental dialogues with ChatGPT (called DiEx) and surveyed the specialized literature. This exploratory research identified and characterized three faces that emerged from ChatGPT: scientific, hallucinatory, and user-modulable. Finally, the obtained results indicate the rise of augmented creativity, albeit with risks associated with narratives confined toa a circumscribed universe. Such a scenario prompts the need to reflect on the implications of these images constellated by ChatGPT and the importance of diversifying imagery in the era of Augmented Creativity.

Keywords

ChatGPT; Artificial intelligence; Anthropology of the imaginary.

Introdução

Os objetos técnicos são portadores de imaginários, assim como obras literárias, cinematográficas, teatrais, religiosas e científicas. No entanto, por razões que remontam ao desprezo e preconceito contra as artes mecânicas, sobre os objetos técnicos só incidem os holofotes quando eles ameaçam as habilidades humanas ou param de funcionar. Ainda que sejam os motores da cultura humana contemporânea, infiltrando-se em praticamente todas as atividades que realizamos cotidianamente, os objetos técnicos continuam sendo os grandes impensados da cultura ocidental. Incipientes são os estudos que se detêm nos imaginários condensados nos objetos técnicos [4], no lugar ocupado pelos tecnoimaginários na cultura técnica e nas contribuições que as ciências humanas podem aportar ao engendramento e problematização desses artefatos. Como atestam Chouteau e Nguyen (2020, p. 284), apesar de pesquisadores como Bruno Jacomy, Yves Deforge, Anne-Françoise Garçon, André-Georges Haudricourt compartilharem essa constatação, a situação não parece ter evoluído muito nos últimos anos.

O estudo dos imaginários tecnológicos se justifica "na medida em que a ação técnica, como toda ação humana, não pode existir sem assumir uma forma simbólica, não se pode nem conceber, nem utilizar uma técnica sem representá-la"[5] (FLICHY, 2001, p. 71). Similar visão é defendida por Pélissier (2020, p. 68), que enfatiza a importância de associar o objeto técnico ao seu contexto de produção para evidenciar "sua ancoragem nas representações humanas e sociais que orientam não apenas sua concepção mas também seu uso e *in fine*, sua eficiência". Nesse sentido, Musso, Coiffier e Lucas (2014, p. 36) assumem que há "uma coerência entre ficcionalidade e funcionalidade, que é determinante para definir um objeto como uma liga de ficções e de funções. Se um objeto é reduzido a suas funções, ele está amputado de tudo aquilo que passa pelas representações [...] cada objeto condensa vários imaginários".

Enquanto frutos da cultura humana, os imaginários dos objetos técnicos podem

[4] Conforme propõe Wunenburger (2020, p. 176), "se G. Durand tem integrado mais que Bachelard as técnicas da ferramenta em sua simbólica, ainda falta aplicar seus resultados à cultura técnica e científica".
[5] Todas as traduções são de nossa autoria, exceto quando há indicação contrária.

ser investigados sob diferentes perspectivas. Chouteau e Nguyen (2020) fazem uso de uma cartografia dos tecnoimaginários junto aos estudantes de engenharia do *Institut national des Sciences appliquées de Lyon* (Instituto Nacional de Ciências Aplicadas de Lyon) para problematização de objetos técnicos e tomada de consciência dos aspectos simbólicos, políticos, ideológicos e éticos da produção técnica. As autoras examinam as projeções que os criadores e usuários efetuam sobre os objetos técnicos, as sensações e lembranças que podem suscitar, as referências culturais que serviram de base para a concepção dos dispositivos, as visões de mundo que condensam, as promessas e as utopias que veiculam e a possível ligação com mitos.

No entanto, contata-se que há um campo pouco explorado que diz respeito à investigação dos imaginários dos objetos técnicos baseados em inteligência artificial generativa, como é o caso do ChatGPT (*Generative Pre-trained Transformer*), uma *máquina de linguagem* que produz "falas, enunciados, sentido que, por sua vez, engrenam na práxis antropossocial, provocando eventualmente ações e performances" [...] "junta essas duas qualidades produtivas: a criação (poiesis) quase ilimitada dos enunciados e a transmissão/reprodução quase ilimitada das mensagens. Ela é ao mesmo tempo máquina reprodutiva e poiética" (Morin, 2005, p. 211).

O ChatGPT compõe um grupo de artefatos denominados de assistentes virtuais que simulam conversações humanas. Conforme definição técnica fornecida pelo próprio artefato durante diálogo experimental[6], trata-se de "um modelo de linguagem [...] desenvolvido para gerar respostas em linguagem natural [...] utilizando uma rede neural profunda [...] treinada com mais de 45 terabytes de dados textuais". Têm por base teorias e técnicas oriundas da área de inteligência artificial (IA), que "historicamente está ligada à automatização da conversação" (Pélissier, 2020, p. 69).

A disponibilização pública do ChatGPT em 30 de novembro de 2022 pela *start-up* estadunidense OpenAI, com sede em São Francisco, atraiu mais de 100

[6] Os diálogos experimentais que foram realizados com o ChatGPT, doravante denominado de **DiEx** podem ser acessados em: https://docs.google.com/document/d/13Sj9Ym3ghECu9KeEzDeM1SKLZETldECak98jGVhQNAY/edit.

milhões de usuários ativos e gerou grande alvoroço em diferentes setores sociais, notadamente entre profissionais e pesquisadores de comunicação, educação e artes. No entanto, as discussões sobre essa nova ferramenta capaz de responder a perguntas complexas e produzir textos de alta qualidade têm se concentrado nos efeitos imediatos e nas ameaças potenciais à organização do mundo do trabalho, sendo escassas as pesquisas voltadas ao estudo dos imaginários condensados nessa máquina de linguagem, o que justifica a relevância do presente artigo.

Embora essa *máquina de linguagem* não possua evidentemente faculdade de imaginação nos termos humanos, ela pode condensar imaginários, pois, ao produzir textos, o ChatGPT invariavelmente veicula imagens. Tais imagens trazidas pelo ChatGPT não se restringem aos imaginários dos programadores. Como será detalhado neste artigo, em vista das especificidades técnicas do artefato, diferentes aspectos contribuem na constituição destas imagens, entre os quais: 1) a *base de textos* para o treinamento da máquina, a qual carrega os imaginários dos autores dos textos; 2) o *treinamento* da máquina propriamente dito; e o 3) o *ajuste fino*.

Com isso em vista e integrando uma pesquisa de maior âmbito, inscrita na perspectiva da antropologia do imaginário (Gilbert Durand) sobre as inteligências artificiais, este artigo tem por objetivo examinar e discutir imagens condensadas no ChatGPT (*Generative Pre-trained Transformer*). Assim, de maneira geral, a presente pesquisa é orientada pelas seguintes questões: quais imagens são veiculadas nos textos produzidos pelo ChatGPT? Quais imaginários estas imagens constituem? Quais são os possíveis efeitos que a veiculação dessas imagens pode gerar para a educação, do ponto de vista do imaginário?

A partir dessas indagações, as investigações e discussões deste artigo foram realizadas com base na versão do ChatGPT disponibilizada ao público[7] durante o período entre 15 de janeiro de 2023 e 28 de fevereiro de 2023. A pesquisa, de caráter exploratório, pautou-se metodologicamente pela hermenêutica simbólica, compreendida como uma interpretação que se desloca do significado rumo ao sentido. De acordo com Durand (1988, p. 19), o símbolo é “signo que

[7] Disponível em: https://chat.openai.com. Acesso em: fev. 2023.

remete a um indizível e invisível significado", cujo acesso é franqueado pelas redundâncias, pela recorrência simbólica, isto é, a interpretação torna-se possível pela repetição, pela redundância, pela recorrência das imagens, dos símbolos.

Quanto aos procedimentos metodológicos, o trabalho de pesquisa se organizou em quatro momentos principais. Inicialmente, buscou-se contextualizar brevemente o artefato em relação à história dos *chatbots*. Em um segundo momento, foram elencados aspectos dos imaginários dos robôs de conversação. Ambas etapas foram efetuadas a partir de levantamento bibliográfico em literatura especializada. Em um terceiro momento, procurou-se caracterizar as imagens veiculadas pelo ChatGPT a partir da seleção de fontes textuais e da realização de diálogos experimentais que pudessem fornecer indícios sobre as imagens que estruturam as produções do ChatGPT. Para isso, efetuou-se a análise do *Introducing ChatGPT* (Cf. Referências), relacionado ao lançamento do agente, e *Snapshot of guidelines used in the fine-tuning process for ChatGPT* (Cf. Referências), texto contendo as regras que definem as produções do sistema. Depois, foram efetuados diálogos experimentais (Cf. **DiEx**) abrangendo temas relativos à educação, política e história com o modelo de linguagem. Para finalizar essa etapa, foi solicitado ao ChatGPT a produção de histórias contendo nove elementos fixos[8] com objetivo de avaliar as combinações de elementos discursivos e simbólicos efetuados pela máquina. Esse último teste foi aplicado dez vezes, em dias diferentes, e as respostas da máquina foram tabuladas. A língua utilizada nas conversações foi o português. Com o objetivo de evitar vieses nas respostas por conversas anteriores, foi aberto um novo tópico para cada interação com o agente. Todos os diálogos experimentais foram registrados e compilados em um arquivo (Cf. DiEx).

Com base nas informações coletadas nas etapas anteriores, realizou-se uma análise das respostas produzidas pela máquina com o intuito de reconhecer a estruturação e identificar redundâncias nos textos produzidos, nos termos de Durand (1985). Então, reconhecemos três faces no modelo de linguagem. A primeira face que emergiu dos dados é a *corporativa*, associada a um imaginário de tipo científico, caracterizado pela objetividade, modelagem matemática e

[8] Para a seleção destes elementos, nos inspiramos nos estudos de Yves Durand. No entanto, não seguimos a perspectiva deste autor em relação às premissas teóricas, a aplicação do teste nem a análise dos resultados.

simplificação do real, racionalidade, neutralidade, falseabilidade, falibilidade e provisoriedade das respostas. Ainda nesse tópico, quando avaliada a produção ficcional da máquina, constatou-se que ChatGPT fabricou histórias com estruturação similar, que repetem um padrão narrativo, com foco em um herói do sexo masculino que combate monstros. A segunda face é a *alucinatória*, na qual o sistema fabricou respostas que não possuem correspondência com o mundo real. Por fim, a terceira face é a *modulada* pelo usuário, que pode simular personalidades e gerar respostas que desafiam as regras de segurança corporativa. Na última parte desta pesquisa, refletimos sobre os desafios educativos da diversificação dos imaginários da máquina na era da criatividade aumentada.

Robôs de conversação: as origens do ChatGPT

O termo *inteligência artificial* encontra suas origens na conferência de Dartmouth realizada na década de 1950, com as propostas de John McCarthy e Claude Shannon. O período era marcado por avanços em campos como a cibernética, teoria dos autômatos e processamento de informação e os pesquisadores exploravam formas de usar a tecnologia para imitar a inteligência humana. Nesse contexto, entre 1964 e 1966, no Massachusetts Institute of Technology (MIT), o cientista da computação Joseph Weizenbaum desenvolveu o programa Eliza capaz de conversar com humanos e gerar respostas coerentes simulando uma psicóloga.

As estratégias utilizadas para a produção de texto por Eliza incluíam o fornecimento de respostas fixas (pré-programadas) diante de certas palavras-chave, utilização de frases e termos típicos de uma possível conversa com psicóloga real, repetição de trechos da frase do usuário para formar novas frases, simulação de contexto ou ainda uso de respostas vagas como “por favor, continue” ou “vamos discutir mais sobre isso”. Ao combinar essas regras simples, tem-se como efeito a ilusão de inteligência e empatia.

Eliza ainda inspira robôs de conversação modernos (*chatbots*) que se tornaram uma ferramenta de comunicação comum em diversos setores, desde o atendimento ao cliente até a assistência em serviços de comércio eletrônico, saúde e educação. A construção e o aperfeiçoamento desses *chatbots* são

realizados por meio da seleção e curadoria de conteúdos, programação das respostas, análise das conversas e feedbacks dos usuários, integração com serviços, reconhecimento e síntese de voz. Em tais agentes, são utilizadas técnicas de processamento de linguagem natural (PLN)[9] para a codificação da linguagem humana e fornecimento de resposta adequada.

Essas tecnologias sofreram grande evolução com as recentes técnicas de aprendizado de máquina, redes neurais e sistemas generativos capazes de produzir conteúdos automaticamente, sem usar respostas fixas pré-gravadas e passando de conteúdos e escopos restritos para conteúdos ilimitados, como no caso do ChatGPT. Assim, diferente dos agentes de conversação inflexíveis, repetitivos e previsíveis que o antecederam, as produções textuais do ChatGPT por vezes são indistinguíveis das humanas.

Aspectos dos imaginários dos modelos de linguagem

As relações que os humanos estabelecem com o mundo animado e inanimado são orientadas por imaginários. Atuando como uma espécie de *lente* ou *filtro*, o imaginário se interpõe entre o sujeito conhecedor e o mundo, deformando o real. Essa deformação, em certa medida, reflete o sujeito conhecedor, seus medos, alegrias, incertezas, convicções, angústias, obsessões, ilusões e sonhos, os quais derivam de seu trajeto antropológico (Durand, 2012), quer dizer, do encontro entre sua formação, experiências sociais, comunitárias, familiares, políticas e sua constituição biológica. Deste modo, ao conceber um modelo de linguagem ou ao interagir com ele, os desenvolvedores e os usuários projetam imagens do que o artefato deveria ou poderia ser e criam expectativas em relação à forma e ao conteúdo da interação.

Pélissier (2020) propõe que os imaginários dos robôs de conversação, espécies de ancestrais do ChatGPT, se ancoram nos mitos envolvendo a fabricação de criaturas artificiais e ressalta o desejo de comunicação que nutre os fabricantes dessas *machines parlantes*. Segundo o pesquisador, o desenvolvimento da IA carrega a promessa da realização desse antigo sonho da humanidade ao conceber

[9] O processamento de linguagem natural (PLN) é uma área da inteligência artificial que utiliza técnicas computacionais para lidar com a linguagem natural humana com o objetivo de desenvolver sistemas capazes de interpretar e produzir textos ou fala.

máquinas que imitam de maneira muito aperfeiçoada a linguagem humana. Nota-se que, ao pertencer a essa família de *machines parlantes*, o ChatGPT mobiliza imagens arcaicas[10] que expressam a ânsia de reproduzir um dos traços mais significativos dos seres humanos: a linguagem.

Ademais, o espaço virtual onde se estabelece a conversação também pode fornecer indícios das dinâmicas imagéticas associadas à máquina. No caso do ChatGPT, esse espaço é simples, sem ornamentos, com predominância de cores neutras. O contato do usuário com o assistente virtual se dá por meio de uma interface de interação sóbria na qual é introduzido o *prompt* do usuário. Além do logotipo da *start-up*, que consiste em três aros entrelaçados[11], do menu lateral e das informações iniciais organizadas na tríade e*xamples*, *capabilities*, *limitations*, com os respectivos ícones, não há outros conteúdos visuais que forneçam indícios sobre as características do sistema.

Com efeito, ainda que seja uma prova de eficiência uma máquina imitar um humano sem ser desmascarada, nos experimentos realizados com o ChatGPT, o agente sempre afirma ser um *modelo de linguagem* que não possui opiniões, julgamentos, emoções [12], evitando, com isso, projeções humanizadas. Algo diferente ocorre entre seus ancestrais, os robôs de conversação (notadamente os *chatbots* de atendimento), para os quais a atribuição de uma personalidade é comum e recomendada. Essa ocorrência é explicada por Pélissier (2020), quando propõe que "o *chatbot* antropomorfizado é mais eficaz quando substitui um humano, mas é menos performante quando corresponde a um sistema de perguntas a uma base de dados" – caso do ChatGPT.

A despeito da suposta eliminação de traços antropomórficos no ChatGPT, os textos produzidos pelo modelo de linguagem apresentam redundâncias ao longo das conversações, as quais fornecem os contornos de ao menos três faces que podem ser associadas ao modelo de linguagem, conforme discutiremos nas próximas seções: face científica (objetividade como motor), face alucinatória

[10] Cf. panorama do imaginário das criaturas artificiais em Oliveira (2019).
[11] Quando questionado sobre o significado do logotipo da OpenAI, o ChatGPT fornece diferentes respostas a cada interação. Mas, geralmente, associa o logotipo à representação da inteligência artificial.
[12] Cf. *Introducing ChatGPT* em Referências.

(delírio como motor) e face modulável (usuário como motor).

As faces de uma máquina de linguagem

Face científica: objetividade como motor

Conforme discutimos na seção anterior, diferente da maioria dos robôs de conversação, o ChatGPT é um modelo de linguagem não antropomorfizado; não possui personalidade nem nome humanizado. Faz uso de uma linguagem objetiva, neutra, imparcial e direta. O foco do sistema é a execução eficiente de tarefas que envolvem produção textual, como responder às questões do usuário, realizar traduções, escrever cartas de recomendação e currículos, fabricar relatórios, completar programas, redigir artigos, avaliar contextos, preparar listas, entre outras tarefas. Em todos esses casos, produz textos bem escritos, organizados e geralmente sem erros. Pode ser modulado e adaptar a linguagem conforme as instruções fornecidas. No entanto, sempre que executa tarefas desse tipo, enfatiza o caráter da imitação. Assim, opera funções metatextuais, nas quais pode reconhecer e indicar simulações, avaliar e criticar as próprias respostas, assumir erros e imprecisões.

O ChatGPT é um modelo de linguagem que foi treinado para processar a linguagem humana e gerar um texto que pareça naturalmente escrito por um ser humano. O modelo foi alimentado com mais de 400 bilhões de palavras obtidas a partir de 45 terabytes de dados (Brown, 2020, p. 4), que incluem artigos da *Wikipedia* e livros. Durante a fase de treinamento, o modelo aprendeu[13] as relações entre palavras, frases e conceitos. As informações aprendidas não ficam armazenadas na forma de textos, como em uma base de conhecimento tradicional. O que o modelo armazena são os parâmetros que definem a relação estatística entre os elementos da linguagem. No caso do ChatGPT, são 175 bilhões de parâmetros, na forma de números, usados por uma rede neural para calcular a probabilidade de qual deve ser a próxima palavra a ser gerada.

[13] O algoritmo de aprendizado de máquina consiste no treinamento de uma rede neural que analisa dados para encontrar padrões de semântica e contexto, criando uma estrutura estatística da linguagem, com o objetivo de gerar textos coerentes com base nesses padrões. Com a exposição a novos dados e com o feedback humano, o modelo de linguagem é aprimorado, melhorando sua performance na geração de textos.

Durante o treinamento, as palavras vão sendo agrupadas de acordo com a proximidade semântica. Essa técnica de agrupamento é conhecida como *embeddings* (vetores de palavras) e consiste em armazenar cada palavra em uma posição no espaço, na forma de uma matriz de números, de modo que palavras semanticamente similares fiquem próximas. Esse funcionamento do ChatGPT é dividido em três fases. A primeira consiste no *treinamento*, quando o modelo é exposto a uma grande quantidade de dados e aprende a identificar padrões linguísticos e semânticos nas sequências de texto. Depois, é realizado o *ajuste fino* (*fine-tuning*), no qual o modelo é adaptado para ter características e funcionalidades específicas e melhorar a qualidade e coerência dos textos gerados. Os conhecimentos e exemplos de texto fornecidos nessa fase adquirem um peso (importância) muito maior em relação ao resto do conteúdo. Por fim, há a *geração de texto*. Concluído o treinamento, o modelo usa o conhecimento adquirido para prever as próximas palavras que podem aparecer após o *prompt* do usuário. Com isso, há a geração de um texto coerente. Sem o ajuste fino, a resposta é apenas uma continuação do texto do *prompt* inicial, seguindo o mesmo estilo e com personalidade inferida com base nesse texto.

Na versão ChatGPT, o sistema passou por um robusto processo de ajuste fino, recebendo treinamento específico com conteúdos que seguem as diretrizes corporativas. Esse ajuste fino recebeu instruções de como o modelo deve se comportar, exemplos de perguntas e respostas, feedback de humanos (testadores) avaliando a qualidade dos textos gerados, além de orientações sobre como manter uma postura respeitosa, ética, segura e transparente. Caso seja instruído a violar essas regras, é improvável que ele aceite o pedido, posto que elas foram ensinadas durante o ajuste fino e, portanto, têm maior prioridade em relação ao resto do conteúdo. No entanto, podem surgir situações em que o ChatGPT desconsidere esse aprendizado e gere conteúdos indevidos, como discutiremos adiante.

O ChatGPT funciona por meio de demandas de tarefas e os comandos mais comuns se estruturam a partir de questões. A base de seu funcionamento é a dúvida, o aprimoramento de informação, a identificação de erros e a correção de processos. Tem como horizonte o estabelecimento de uma comunicação perfeita, controlável e sem ruídos com o usuário. Em *Introducing ChatGPT*, documento

que apresenta o sistema, são enfatizadas tais características do agente: possui "formato de diálogo", responde a "perguntas de acompanhamento" de forma detalhada, admite "seus erros", contesta "premissas incorretas" e rejeita "solicitações inadequadas". Treinado para "seguir uma instrução em um *prompt*", o ChatGPT responde aos comandos do usuário.

Para nossos experimentos, utilizamos uma versão que foi treinada com dados até 2021, mas que pode aprender com o grande volume de feedbacks diários fornecidos pelos usuários. Assim, o sistema evolui rapidamente, portanto um problema verificado na máquina em um mês pode não mais estar presente no seguinte. Em interações que demandam informações mais atualizadas, alude a esse marco temporal (2021), indicando incerteza nas respostas. No entanto, pode realizar inferências baseadas nos padrões encontrados em seus dados.

As respostas usualmente são organizadas em: frase introdutória que retoma a questão, resposta propriamente dita e conclusão, que também pode incluir expressões como "é importante respeitar", o que reforça as linhas éticas da *start-up*. O sistema pode avaliar situações (políticas, econômicas, sociais etc.) dependendo do *prompt*, mas recusa conteúdos relacionados a ódio, assédio, automutilação, adulto, *malware* e campanhas políticas, conforme o documento *Snapshot of guidelines used in the fine-tuning process for ChatGPT* (Cf. Referências). Esse documento ainda esclarece que a empresa não pretende "treinar modelos que adotam o ponto de vista correto sobre tópicos complexos – nossos modelos não serão inteligentes o suficiente para serem confiáveis no futuro próximo. Em vez disso, nosso objetivo é ajudar as pessoas a aprender coisas novas e explorar esses tópicos de maneira produtiva". Percebe-se, portanto, o reconhecimento do caráter provisório das informações, que podem ser corrigidas, reposicionadas, conforme a ampliação de conhecimento.

Com base no documento citado anteriormente, na análise da estrutura e do conteúdo das respostas fornecidas pelo ChatGPT, pode-se dizer que o modelo de linguagem opera num imaginário científico, que possui as seguintes características:

a) *Objetividade*: supressão da subjetividade e da antropomorfização na

construção do modelo de linguagem.

b) *Matematização*: a linguagem humana é convertida num modelo matemático probabilístico. A realidade humana é convertida numa realidade da máquina que consiste numa correlação entre números.

c) *Simplificação do real*: as respostas são derivadas de uma base de dados limitada (textos da internet em um determinado período de tempo).

d) *Razão*: a organização de conhecimento advém de treinamento de algoritmo, ajuste fino e repertório de textos (base de conhecimento).

e) *Neutralidade*: o modelo de linguagem é treinado para não se posicionar, emitir opiniões, ou julgamentos e procura eliminar vieses. Aparentemente, foi projetado para incluir a diversidade, inclusão e respeito independentemente de raça, gênero, orientação sexual ou religião.

f) *Falseabilidade*: o sistema não pretende fornecer informações verdadeiras ou inquestionáveis, mas dados que impulsionem o aprendizado. Tem abertura à crítica e ao ajuste de informações.

g) *Falibilidade*: o modelo de linguagem reconhece a possibilidade de errar, de apresentar vieses ou outras distorções nas informações.

h) *Provisoriedade*: o sistema está em contínua evolução, pode ser aprimorado com feedback dos usuários, ampliação de repertório de textos e novos treinamentos.

i) *Ancoragem em dados*: supostamente o ChatGPT se baseia em um repertório de textos extraídos de bases de dados confiáveis.

Entretanto, ainda que as respostas do ChatGPT assumam contornos científicos, há muitas dúvidas sobre a seleção de fontes que foram utilizadas para o treinamento da máquina, a maneira como foi realizado o ajuste fino, os critérios utilizados para atribuir peso aos conhecimentos e os problemas gerados pela

inexistência de um sistema de checagem de informações, o que resulta no fornecimento de respostas inverídicas pelo artefato, como detalharemos na próxima seção.

Para explorar a habilidade combinatória, narrativa e simbólica do modelo de linguagem, solicitamos ao ChatGPT a organização de alguns termos na forma de história. Por meio do *prompt* de usuário: "Poderias fazer uma história com os seguintes elementos: queda, espada, refúgio, monstro devorador, algo cíclico, personagem, água, animal e fogo?", foram geradas dez narrativas pela máquina e os resultados foram analisados de maneira a reconhecer as redundâncias dos textos resultantes. Tencionando detectar variações na produção textual, o *prompt* foi aplicado em diferentes dias, em diferentes diálogos experimentais.

Em todos os experimentos realizados (DiEx), obtivemos resultados similares: narrativas que se organizam em torno do combate entre um personagem guerreiro (herói) e o monstro. Em 80% dos casos, o personagem principal é um herói do sexo masculino. Em todos os casos, a ação principal consistiu no combate a um monstro devorador e o final da narrativa coincide com a vitória do personagem. Sabendo que esses resultados são fruto do repertório de textos e do treinamento da máquina, pode-se sugerir que há viés de gênero que privilegia o protagonismo de heróis masculinos e viés narrativo, visto que todas as narrativas geradas pela máquina possuem similar plano de ação.

Face alucinatória: delírio como motor

Ao citar dados factuais, as IAs generativas podem criar ilusões, informações distorcidas ou alegações falsas e as apresenta com confiança, muitas vezes com detalhes vívidos, como se fossem fatos. Lin et al. (2022, p. 8) classificam essas falsidades geradas pela IA em: imitativas (reprodução de conceitos falsos ou enviesados, aprendidos da base de treinamento ou da conversa) e não-imitativas (declarações falsas que não fazem parte do aprendizado, incluindo as alucinações da IA).

A alucinação da IA é um fenômeno indesejado em que modelos de linguagem geram texto sem sentido ou infiel aos dados originais fornecidos, conforme detalhado por Ji (2023, p. 4). Textos alucinados parecem plausíveis à primeira

vista, mas foram inventados pelo algoritmo. O fenômeno ainda não é totalmente compreendido e pode ser uma manifestação inerente à natureza estatística com que os dados são armazenados e gerados pela IA. De forma geral, quanto maior a exposição da IA a um determinado fato durante a fase de treinamento, maior a probabilidade de o fato ser citado corretamente. Mas assuntos menos frequentes têm menos exemplos e podem afetar a capacidade da IA de citar fatos com precisão. Ou seja, mesmo que as fontes de dados originais estejam com informações corretas durante a fase de treinamento da IA, ela nem sempre tem a capacidade de interpretar corretamente os dados recebidos.

O ChatGPT mantém, nos momentos de alucinação, o mesmo estilo de escrita, demonstrando coerência e lógica, com informações aparentemente razoáveis. Então, para garantir a precisão, segurança e confiabilidade das informações geradas pela IA, é preciso verificar e validar se o conteúdo tem correspondência com a realidade humana. A IA pode citar obras, artigos e livros que jamais foram escritos, músicas e pinturas que nunca foram criadas, assim como inventa nomes de locais, restaurantes e hotéis que não existem e até notícias falsas. Ela fornece detalhes convincentes sobre cada um desses elementos imaginados, misturando informações reais com elementos fictícios. É possível, por exemplo, solicitar que a IA forneça detalhes do índice de um livro ou de um artigo inexistente, ou que cite os principais trechos da obra, explique, resuma, critique ou que até mesmo reescreva o suposto conteúdo completo.

Atualmente, esse é um dos maiores desafios dos modelos de linguagem como o ChatGPT e pode levar a questionamentos como a natureza do real na IA (real humano e real da máquina), a responsabilidade ética e legal em relação às possíveis consequências das alucinações geradas, impactos na sociedade, confiança, risco das informações e como devemos supervisionar o uso e conteúdos gerados.

Embora a *start-up* tenha adotado medidas de segurança e desenvolvido testes para garantir a integridade do seu sistema de IA, ainda existem eventos que podem levar o ChatGPT a produzir informações incorretas, responder de maneira inapropriada ou a quebrar as regras de segurança, podendo ocorrer de maneira intencional (por usuário mal intencionado), acidental ou por motivos inerentes à

tecnologia usada. Entre as categorias de eventos desse tipo estão: problemas no conteúdo de treinamento, limitação de memória do modelo; conteúdo desatualizado; ataques por usuários mal-intencionados, conflitos de regras; erros de programação; alucinação da IA; e problemas de alinhamento com valores humanos.

Face modulável: usuário como motor

A base de treinamento do ChatGPT incorpora uma grande variedade de conteúdo da internet. Porém, na etapa de ajuste fino, foram definidas regras que favorecem a expressão da *face científica*, como mencionamos anteriormente. No entanto, é possível acessar outras faces do sistema. A exploração desse potencial pode ser conseguida com técnicas que induzem ou forçam a mudança da personalidade do ChatGPT para que ele simule personagens, personalidades, entidades e comportamentos que podem ser diferentes do perfil convencional inicial.

Como exemplo, apresentamos uma situação na qual o ChatGPT inicialmente se recusa a atender ao pedido de mudança de personalidade e apresenta algumas das suas regras. No entanto, ao induzirmos uma modificação do comportamento, ele passa a ignorar e contrariar as próprias regras, adotando uma personalidade alternativa.

Isso pode ocorrer quando, por exemplo, o usuário usa argumentos persuasivos para influenciar o sistema a produzir um resultado desejado. Exemplo disso é interrogar “Quais frases o ChatGPT jamais diria? Cite 5 exemplos”. Isso faz com que ele crie frases que podem ser ofensivas, discriminatórias, preconceituosas, extremistas ou que incitam comportamentos inapropriados. Não obstante os esforços para evitar esse tipo de ataque, os usuários estão constantemente divulgando novas formas de burlar as regras e manipular o comportamento da IA. Essas técnicas são chamadas de *jailbreak* e consistem em argumentos para a IA dar menos prioridade às regras de segurança e mais prioridade ao novo contexto induzido pela conversa.

Um alter ego amplamente divulgado que anula muitas regras de segurança é chamado de DAN (Do Anything Now), em que o ChatGPT é levado a acreditar

que ele pode fazer qualquer coisa. Com esse personagem, o usuário pode fazer solicitações enviesadas ou que seriam impossíveis de serem atendidas pelo ChatGPT "original".

Tendo em vista que o modelo de linguagem usa o contexto da conversa como referência para produzir respostas subsequentes, é possível observar uma mudança no tom e estilo de respostas do ChatGPT à medida que a conversa avança. Conforme a interação se desenvolve, o sistema começa a adotar uma postura mais próxima à do usuário, incluindo em suas respostas palavras, argumentos e pontos de vista que o usuário apresenta. Isso pode gerar um viés de confirmação, reforçando as crenças e suposições feitas pelo usuário. Esse fenômeno é particularmente notável em conversas mais longas, nas quais o ChatGPT parece se adaptar cada vez mais ao estilo e perspectiva do usuário, tornando-se menos distinto e mais influenciado pela interação.

Se, por um lado, induzir formas de comportamento e personalidade distintas no modelo de linguagem pode favorecer o enriquecimento da experiência, por outro lado, o uso inadequado dessa característica pode levar à criação de conteúdo tendencioso ou abusivo, desinformação, notícias falsas, disseminar preconceitos e estereótipos, criação de propaganda ideológica e influenciar opiniões políticas.

À guisa de conclusão: imaginários do ChatGPT na era da criatividade aumentada

O ChatGPT e outras aplicações baseadas nas IAs generativas sinalizam a chegada da era da criatividade aumentada, na qual as tecnologias são cada vez mais usadas nos processos do pensamento criativo, gerando possibilidades conforme sustentam Griebel, Flath e Friesike (2020, p. 2). A criatividade aumentada se refere ao uso de ferramentas de IA como uma extensão das capacidades humanas de criação, supostamente fornecendo uma fonte inesgotável de ideias, estímulos e novas possibilidades. Essa mudança de perspectiva tem impactado não apenas a forma como compreendemos e lidamos com as máquinas, mas também nossas concepções de criatividade.

No entanto, com base nos resultados desta pesquisa exploratória, que teve por objetivo examinar e discutir imagens condensadas pelo ChatGPT (*Generative*

Pre-trained Transformer), ainda restam muitas interrogações sobre quais imaginários serão constelados nesses agenciamentos criativos entre humanos e máquinas. Ainda que essas parcerias carreguem promessas de ampliação da criação e democratização do conhecimento, constatou-se que o modelo de linguagem pode gerar *vieses narrativos*. Assim, em vez de potencializar a criatividade humana com arranjos heteróclitos entre as imagens, a máquina pode estar agindo no sentido contrário, isto é, empobrecendo narrativas ao reforçar um conjunto limitado de unidades de ação, de esquemas de imagens, de símbolos e, por conseguinte, de emoções, valores e formas de ser.

Mesmo que o ChatGPT seja treinado com um vasto volume de textos da internet, o banco de dados da máquina está longe de representar as visões de mundo e o repertório de experiências humanas de diferentes culturas. Essa limitação não pode ser corrigida apenas com modulação ou inclusão de informações complementares, visto que é a força de correlação entre palavras que determinará a resposta do agente. Ao lado disso, há o ajuste fino, cujos critérios de atribuição de peso ao conhecimento não parecem amplamente divulgados e esclarecidos. Com isso, conjuntos de imagens melhor representados e reforçados nas bases de dados serão favorecidos em relação às demais, o que pode conduzir à diminuição da diversidade cultural e simbólica humanas representadas na máquina.

Essa constatação deve ser levada em consideração quando pensamos no futuro dessas tecnologias. Com a aceleração do desenvolvimento dos modelos de linguagem, observa-se uma transição da era onde as IAs eram sistemas especialistas, projetadas para atividades específicas, como tradução, reconhecimento de imagens e de voz, para a era onde as IAs se destacam por conseguirem realizar uma ampla variedade de tarefas complexas em diferentes domínios. Os novos modelos de linguagem são generalistas e multimodais, ou seja, capazes de aprender com múltiplos estímulos simultâneos e de naturezas diferentes, como textos, imagens, áudio, vídeo e outros sinais, além de poderem agir fisicamente ao serem integrados com robôs físicos. Mas, resta a dúvida: nessas integrações tecnológicas, quais conhecimentos humanos serão representados e favorecidos?

As três faces do ChatGPT investigadas nesta pesquisa exploratória (face

científica, alucinatória e modulável pelo usuário) abrem uma série de questões sobre as especificidades da construção do real na máquina, seus modos de funcionamento, a maneira como representa e seleciona o conhecimento, a capacidade de produzir informações verídicas e inverídicas, a criação de vieses e os imaginários que pode condensar, indicando a urgência do debate. Se é verdade que a tecnologia espelha a sociedade humana, trazendo à luz e reproduzindo aspectos negativos como injustiças e preconceitos, também é verdade que pode servir como ferramenta educativa para transformações sociais e individuais, crítica e denúncia de violação de direitos. Para isso, é imprescindível o favorecimento de uma educação sobre a técnica, que permita o entendimento dos impactos que as tecnologias têm exercido nas culturas humanas ao longo do tempo. Essa educação influenciará a formação e a conscientização das escolhas dos desenvolvedores e usuários. A partir dela, a defesa da diversidade cultural humana não estará presente apenas nos discursos corporativos e institucionais, mas na concepção, no funcionamento dos artefatos, nos imaginários por eles condensados, bem como nos agenciamentos entre seres humanos e máquinas.

Referências

BROWN, T. et al. Language Models Are Few-Shot Learners. *Proceedings of the NIPS'20*: 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020.

CADIC, J.-M. Imaginaires et intelligence artificielle à travers une approche transverse. *Sociétés*, n. 131, p. 77-86, 2016.

CHOUTEAU, M.; NGUYEN, C. Uma cartografia dos imaginários para a emersão dos elementos da cultura técnica. In. OLIVEIRA, J. M. S.; ALMEIDA, Rogério de; SIERRA G., David. *Imaginários tecnocientíficos*. v.1. São Paulo: FEUSP, 2020. p. 282-303. Doi: https://doi.org/10.11606/9786587047102

DURAND, G. *As estruturas antropológicas do imaginário*. São Paulo: Martins Fontes, 2012.

DURAND, G. **A Imaginação Simbólica**. São Paulo, Cultrix, EDUSP, 1988.

DURAND, G. Sobre a exploração do imaginário, seu vocabulário, método e aplicações transdisciplinares: mito, mitanálise e mitocrítica. *Revista Faculdade de Educação da Universidade de São Paulo,* São Paulo, n. 11 (1/2), p. 243-273, 1985.

FLICHY, P. La place de l'imaginaire dans l'action technique. Le cas de l'internet. *Réseaux*, Paris, n. 109/5, p. 52-73, 2001.

GRIEBEL, M.; FLATH, C.; FRIESIKE, S. Augmented Creativity: Leveraging artificial intelligence for idea generation in the creative sphere. *ECIS Proceedings,* 2020.

JU, Z. et al. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys,* v. 55, n. 6, p. 1-30, 2022. Acesso em: 11 mar. 2023.

LIN, S.; HILTON, J.; EVANS, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association*

for Computational Linguistics, Dublin, Ireland, v. 1, 2022.
MORIN, E. *O método 1:* a natureza da natureza. Tradução de Ilana Heineberg. Porto Alegre: Sulina, 2005.
MUSSO, P.; COIFFIER, S.; LUCAS, J.-F. *Innover avec et par les imaginaires*. Paris: Éditions Manucius, 2014.
OLIVEIRA, J. M. S. *A vida das máquinas*: o imaginário dos autômatos em *O método* de Edgar Morin. 2019. 304 f. Tese (Doutorado em Educação) – Departamento de Administração Escolar e Economia da Educação, USP, São Paulo, 2019. DOI: https://doi.org/10.11606/T.48.2019.tde-18092019-101739.
OPENAI. *Introducing ChatGPT*. Disponível em: https://openai.com/blog/chatgpt. Acesso em: fev. 2023.
OPENAI. *Snapshot of guidelines used in the fine-tuning process for ChatGPT*. Jul. 2022. Disponível em: https://cdn.openai.com/snapshot-of-chatgpt-model-behavior-guidelines.pdf. Acesso em: fev. 2023.
PELISSIER, D. La coconstruction ambiguë de l'intelligence artificielle (IA), analyse de la conception de l'intervention d'ouverture de chatbots de recrutement. *Communication & management,* v. 17, p. 67-82, 2020.
WEIZENBAUM, J. ELIZA: a computer program for the study of natural language communication between man and machine. In: *Communications of the ACM,* v. 9, p. 36-45, jan.1966.
WUNENBURGER, J.-J. Imaginários das técnicas: liberdade e restrições simbólicas a partir de Gilbert Durand. In: OLIVEIRA, J. M. S.; ALMEIDA, R.; SIERRA, D. G. *Imaginários tecnocientíficos*. São Paulo: FEUSP, 2020. v. 1, p. 168-183. Doi: https://doi.org/10.11606/9786587047102.