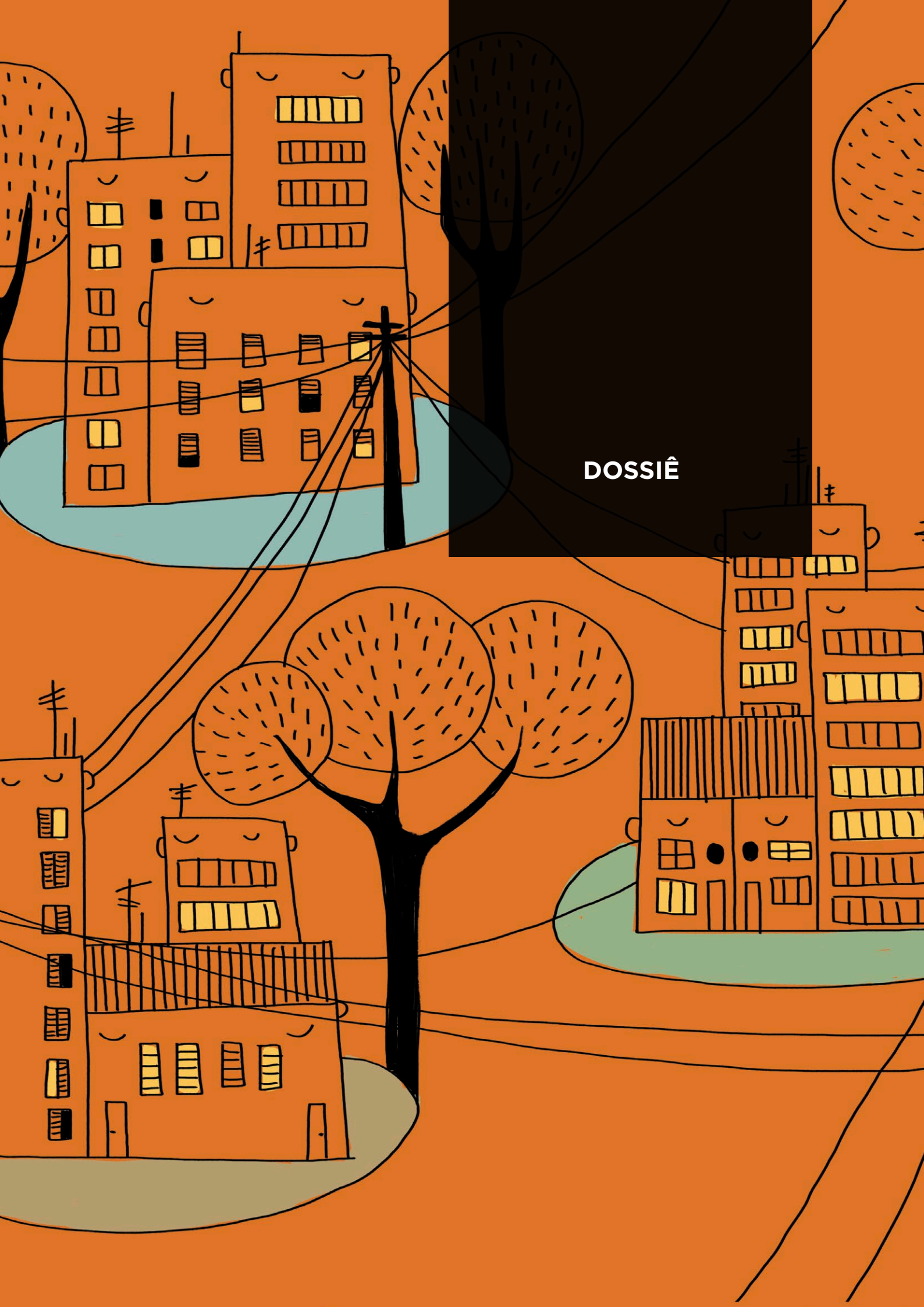
DOSSIÊ

A CARACTERÍSTICA GENERATIVA DAS INTELIGÊNCIAS ARTIFICIAIS E SEUS IMPACTOS SOBRE A CRIATIVIDADE

Sergio José Venancio Júnior[1]

RESUMO

Modelos de inteligência artificial (IA) para geração de texto e imagem, como ChatGPT, Dall-E e Stable Diffusion, são tecnologias que têm sido alvo de processos jurídicos por violação de direitos autorais. A promessa de que essas ferramentas são criativas e de que geram artefatos inéditos entra em controvérsia com tais acontecimentos. O presente texto propõe uma discussão sobre uma compreensão conceitual dos principais processos algorítmicos que explicam o funcionamento de modelos discriminativos e generativos, enquanto busca, também, esclarecimentos sobre as motivações dos mencionados processos litigiosos e sobre as noções de criatividade e da cocriação humano-máquina.

Palavras-chave: Inteligência artificial. Criatividade. Generativo. Arte. ChatGPT.

ABSTRACT

Artificial intelligence models for text and image generation such as ChatGPT, Dall-E and Stable Diffusion are technologies that have been subject to litigation due to copyright infringement. The promise that these tools are creative and generate novel artifacts is contested by such events. This work proposes a conceptual understanding of the main algorithmic processes that explain the operation of these discriminative and generative models, while seeking to clarify the motivations behind the mentioned litigation processes and behind notions of human-machine creativity and co-creation.

Keywords: Artificial intelligence. Creativity. Generative. Art. ChatGPT.

1 Doutorando em Artes Visuais pela Escola de Comunicações e Artes (ECA) da Universidade de São Paulo (USP). Mestre em Artes Visuais pela mesma instituição. Bacharel em Artes Visuais e em Ciência da Computação pela Universidade Estadual de Campinas (Unicamp). Professor de Design e Computação Gráfica no Instituto de Tecnologia e Liderança (Inteli). Pesquisador na área de Artes Visuais e Inteligências Artificiais. E-mail: svenancio@gmail.com.

INTRODUÇÃO

Em dezembro de 2023, o jornal *The New York Times* abriu uma ação no Tribunal Federal de Manhattan contra as empresas OpenAI e Microsoft, mantenedoras do *software* ChatGPT, alegando violação de direitos autorais sobre artigos produzidos por tal veículo de comunicação (Grynbaum; Mac, 2023). O ChatGPT[2] representa hoje o estado da arte dos *softwares* de inteligência artificial (IA): um *chatbot* capaz de manter conversas plausíveis com o usuário e gerar, entre outras coisas, textos supostamente inéditos, com vários parágrafos, com gramática e ortografia corretas, em diversos idiomas, sobre os mais variados temas e sob diversos estilos de redação – tudo isso a partir de pedidos simples, chamados *prompts*, que o usuário pode escrever em uma interface muito similar à de aplicativos de trocas de mensagens, como o WhatsApp.

Para conseguir gerar esse tipo de conteúdo supostamente inédito e espontâneo, o ChatGPT é "treinado" com bilhões de exemplos de textos, dos mais variados contextos, encontrados publicamente na *internet*. Embora não seja recente, o ChatGPT só se tornou oficialmente acessível ao público como um serviço ao final de 2022, e desde então tem sido utilizado nas mais diversas áreas para automatização de tarefas de criação textual. O *software* possui hoje uma versão gratuita limitada e uma versão de assinatura paga, que permite acesso ao modelo gerador mais desenvolvido e atualizado, o qual responde analisando e gerando não somente textos, mas também imagens e código de programação[3].

Esse processo judicial do *The New York Times* é precedido por ações com acusações similares a outros sistemas de IA nos Estados Unidos: a empresa advocatícia californiana Joseph Saveri Firm, por exemplo, abriu ações coletivas contra as empresas mantenedoras do sistema GitHub Copilot em 2022 (Butterick, 2022); e contra as que comercializam o modelo Stable Diffusion em 2023 (Butterick, 2023).

O GitHub Copilot[4] se trata de um sistema pago, capaz de gerar códigos de programação rapidamente, em diferentes linguagens, a partir de *prompts*[5]. Esse sistema vem automatizando a escrita de códigos para programadores no mundo todo. A citada ação coletiva da classe de desenvolvedores de *software* contra o GitHub e a Microsoft, mantenedoras do Copilot,

2 Ver: https://openai.com/chatgpt. Acesso em: 5 jan. 2024.

3 Alguns exemplos de *prompt* no ChatGPT seriam: "Escreva-me um texto de apresentação para o meu currículo, ressaltando não apenas minhas habilidades técnicas, mas também minhas habilidades socioemocionais" ou "Escreva um artigo com cinco parágrafos, dissertando sobre a guerra entre Israel e Hamas".

4 Ver: https://github.com/features/copilot. Acesso em: 5 jan. 2024.

5 Um exemplo de *prompt* no Copilot seria: "Escreva uma função que concatene duas *strings*", e o código é escrito automaticamente no editor de código Visual Studio Code.

alega que o *software* é "um produto de IA que depende de uma pirataria sem precedentes de *software* de código aberto" (Butterick, 2022, tradução nossa) e que "Ao treinarem seus sistemas de IA em repositórios públicos do GitHub, [...] os réus violaram os direitos legais de um grande número de criadores que publicaram código ou outro trabalho sob certas licenças de código aberto no Github" (Butterick, 2022, tradução nossa).

O GitHub é um símbolo do movimento *open source* (código aberto), que, conforme suas licenças de uso mais comuns, permite o reaproveitamento de código público em qualquer outro projeto, contanto que se atribua créditos aos autores e que se produza código derivado também *open source*. Sob o controle da Microsoft desde 2018[6], o GitHub se tornou uma plataforma integrada aos ambientes de desenvolvimento de *software,* como o Visual Studio Code, em que se instala o Copilot como um *plugin* que auxilia o programador a automatizar a escrita de código. A suposta pirataria que o Copilot faz de milhões de repositórios de código aberto se daria a partir do momento em que, segundo a documentação do processo, é possível reconhecer trechos de código gerados muito similares a fontes com autoria definida dentro do próprio GitHub[7].

Por sua vez, o processo contra as empresas que utilizam a tecnologia Stable Diffusion para fins comerciais mobiliza a classe de artistas visuais em uma ação coletiva, alegando que o sistema é "uma ferramenta de colagem do século 21 que remixa obras protegidas por direitos autorais de milhões de artistas cujos trabalhos foram usados como dados de treinamento" (Butterick, 2023, tradução nossa).

O Stable Diffusion, assim como as também famosas tecnologias Midjourney[8] e Dall-E[9], é um conjunto de modelos abertos de geração de imagens e vídeos inéditos feitos a partir de *prompts* de texto (*text-to-image*)[10] ou de outras imagens (*image-to-image*). Essa tecnologia é desenvolvida e mantida pela empresa Stability AI[11], uma das acusadas na ação coletiva dos artistas. Os modelos são treinados a partir de um conjunto de dados (*dataset*) público de mais de cinco bilhões de pares de texto e imagem,

6 Ver: https://news.microsoft.com/2018/06/04/microsoft-to-acquire-github-for-7-5-billion/. Acesso em: 5 jan. 2024.
7 Segundo o documento de queixa, o Copilot foi treinado em trechos de código protegidos por direitos autorais e os reproduz sem seguir as licenças de código aberto. Ver: https://githubcopilotlitigation.com/pdf/06823/1-0-github_complaint.pdf, páginas 12-32. Acesso em: 7 jan. 2024.
8 Ver: https://www.midjourney.com/home. Acesso em: 13 abr. 2024.
9 Ver: https://openai.com/dall-e-3. Acesso em: 13 abr. 2024.
10 Um exemplo de *prompt text-to-image* para Stable Diffusion e outros modelos geradores de imagem como Dall-E e Midjourney seria: "Crie uma cidade submersa, com ciclistas andando em avenidas e animais marinhos nadando em meio a prédios".
11 Ver: https://stability.ai/. Acesso em: 13 abr. 2024.

chamado LAION-5B[12]. Por serem abertos, os códigos, *datasets* e modelos treinados pela empresa podem ser acessados por qualquer pessoa. Para aqueles que não querem se ocupar de processar o modelo em suas próprias máquinas, a Stability AI hospeda seus modelos geradores de imagens e vídeos na nuvem[13] e vende créditos de processamento para uso deles, além de planos de assinatura que garantem ao usuário direitos comerciais sobre as imagens que ele produz.

A referida ação coletiva dos artistas acusa as empresas envolvidas de utilizarem grande quantidade de imagens protegidas por direitos autorais para treinar os modelos e comercializá-los posteriormente, sem o consentimento desses profissionais. Alega ainda que os produtos de tais modelos têm um grande potencial de plágio, uma vez que podem se assemelhar muito aos trabalhos originais dos artistas. A ação hoje teria base em evidências como as levantadas pelo psicólogo e cientista cognitivo Gary Marcus e pelo artista Reid Southen (2024). Os pesquisadores realizaram experimentos com *prompts*, chegando a resultados que claramente evidenciam plágios de trabalhos autorais, especialmente daqueles que fazem parte de um imaginário popular global.

Se observarmos os discursos comerciais em torno de tais tecnologias, teremos promessas de tradução de pedidos simples de texto para qualquer conteúdo textual e audiovisual inédito, o que garantiria produtividade para criadores de conteúdo. Veremos também discursos de "criatividade" das máquinas, realimentando debates que remontam a René Descartes (2017, p. 95-96), Ada Lovelace (1842 apud Turing, 1950, p. 450) e Alan Turing (1950), chegando a um ponto em que a ciência supostamente provou que é possível termos máquinas que criam[14].

Tem sido fascinante, suspeita e amedrontadora a ideia de que as máquinas podem aprender, passar a cumprir tarefas cognitivas e entregar artefatos criativos tão bem ou melhor que muitos humanos. Entretanto, dados os atuais movimentos litigiosos entre criadores e empresas detentoras de modelos de IA, podemos nos questionar se o foco do debate deveria ser esse aspecto de produtividade, ou essa suposta criatividade

12 Ver: https://laion.ai/. Acesso em: 13 abr. 2024.

13 "*Software* em nuvem" se refere à hospedagem de *softwares* em ambientes computacionais remotos, de modo que não é necessário que o usuário armazene tais *softwares* ou execute suas funcionalidades em sua própria máquina. Todo o processamento é feito em máquinas remotas e os resultados são enviados via *internet*. No contexto dos grandes modelos de inteligência artificial de acesso público, o uso de grandes *datacenters* em nuvem para hospedagem e processamento dos modelos tem sido imperativo.

14 O debate sobre "podem as máquinas criar?" foi amplamente discutido por nós anteriormente. Ver Venancio Júnior (2019a, 2019b).

das máquinas e a obsolescência humana em tarefas cognitivas e poéticas. Assim, propomos um entendimento sobre o que de fato acontece com essas novas tecnologias de inteligência artificial, ditas generativas, para esclarecermos estes atuais momentos e seus possíveis desdobramentos e impactos.

Nossa proposta neste texto é explicar alguns fundamentos técnicos e críticos da inteligência artificial e do aprendizado de máquina (*machine learning*), permitindo-nos compreender as causas não apenas dos mencionados processos jurídicos, mas, antes disso, os diversos outros impactos da IA na sociedade contemporânea. Acreditamos que uma compreensão minimamente conceitual e crítica sobre tecnologias digitais e seus algoritmos amplia noções de literacia digital e de cidadania global. Almejamos que um número cada vez maior de pessoas consiga compreender causas e efeitos das complexas tecnologias digitais com as quais convivem e aprendam a ter maior controle sobre elas. Nesse sentido, buscamos oferecer uma abordagem didática de compreensão dessa tecnologia sob duas perspectivas complementares, relacionadas a dois verbos, duas ações: julgar e gerar.

Quando se considera que as IAs são supostamente capazes de realizar julgamento e geração, é possível imaginar um ciclo criativo próximo ao de um humano: geramos algo, julgamos o que geramos, e então geramos novamente de forma melhorada, para novos ciclos de julgamentos e gerações, até que finalmente o criador julgue que algo está satisfatório. Essa possível decomposição de um processo criativo nos norteará nas discussões aqui propostas, em que trataremos criticamente do tema da cocriação entre artistas e máquinas.

INTELIGÊNCIAS ARTIFICIAIS DISCRIMINATIVAS

Inteligência Artificial é uma área da Ciência da Computação que busca representar processos cognitivos através de modelos matemáticos. Seus produtos são *softwares* que simulam processos de uso de linguagem natural, capacidade de análise e abstração, visão, audição, percepção, aprendizado etc. O termo *inteligência artificial* remonta a 1956 e foi supostamente cunhado pelo cientista da computação estadunidense John McCarthy, ao convidar alguns cientistas para um grupo de trabalho na Universidade de Dartmouth. Conforme a proposta:

> Propomos que um estudo de 2 meses e com 10 homens sobre inteligência artificial seja conduzido durante o verão de 1956 na Dartmouth College em Hanover, New Hampshire. O estudo visa prosseguir com base

> na conjectura de que todo aspecto do aprendizado ou qualquer outra característica da inteligência pode em princípio ser tão precisamente descrito de forma que uma máquina poderia ser criada para simulá-lo. Uma tentativa será feita para encontrar como fazer máquinas usarem linguagem, formar abstrações e conceitos, resolver tipos de problemas hoje reservados a humanos e melhorar a elas mesmas. Nós acreditamos que um avanço significativo pode ser feito em um ou mais desses problemas se um grupo cuidadosamente selecionado de cientistas trabalhar em conjunto no assunto durante um verão (Russell; Norvig, 2010, p. 17, tradução nossa).

Não é necessário dizer que a duração desse grupo de trabalho foi insuficiente para conseguir resultados satisfatórios, dado que muitos dos problemas tratados por ele ainda hoje estão em aberto. Mas é interessante notar que a área de Inteligência Artificial, desde seus primórdios, busca aproximações com processos cognitivos de percepção, análise, síntese, tomada de decisão e aprendizado. Segundo Russell e Norvig (2020)[15], tal área da computação se baseou, ao longo de sua história, em duas diferentes abordagens, duas questões: "Estamos preocupados com pensamento [racional] ou [simular] comportamento? Queremos criar modelos que se assemelham a humanos ou tentar alcançar resultados otimizados?" (Russell; Norvig, 2020, p. 34, tradução nossa). De acordo com os autores, historicamente, essa área obteve maior êxito ao se aproximar de modelos de racionalização, que buscavam a melhor ação possível em determinadas situações passíveis de descrição objetiva. Entretanto, os autores propõem dois refinamentos contemporâneos a tais modelos: que as escolhas de ações racionais a serem modeladas sejam também limitadas pelas próprias intratabilidades computacionais; e que a ideia de uma máquina que busca a melhor decisão com base em objetivos bem definidos seja substituída pela ideia de uma máquina que tome a melhor decisão de acordo com objetivos que beneficiem humanos, ainda que tais objetivos sejam incertos (Russell; Norvig, 2020, p. 34).

Dentro da área de IA, modelos que simulam o aprendizado ganharam muita tração nos últimos anos. O chamado *aprendizado de máquina* (*machine learning*) se tornou a técnica mais proeminente para a modelagem de tarefas cognitivas. Em sua essência, o *aprendizado de máquina* propõe modelos estatísticos que se baseiam em conjuntos de dados – os chamados *datasets* de treinamento – para "aprender" padrões numéricos e utilizá-los para discriminar outros dados que não estavam previstos nesse conjunto de treinamento. Em outras palavras, dado um conjunto de dados digitalizados – que podem ser tabelas, textos, imagens, sons, etc. –, um modelo

15 Stuart Russell e Peter Norvig são autores do livro citado, que por sua vez é obra amplamente utilizada em disciplinas de IA nas universidades do mundo todo.

estatístico reconhece padrões numéricos nesse conjunto e os utiliza para classificar novos conjuntos de dados. Tal modelo passa a ser capaz de minimamente discriminar se determinado dado se assemelha ou não a um padrão aprendido anteriormente, e essa é a função primordial de um modelo de aprendizado de máquina: simular a capacidade de discriminar.

Para ilustrar a ideia de modelos de IA discriminativos, consideremos um conjunto de dados com milhares de fotos de cães e gatos domésticos das mais diversas raças, com características físicas diferentes, em diferentes posições, situações, iluminações. Cada uma dessas fotos deve ser rotulada com "cão" ou "gato", simplesmente. Esse será nosso *dataset* de treinamento de um modelo discriminativo. Um algoritmo de aprendizado de máquina[16] processará todas essas fotos, e elas nada mais são do que conjuntos ordenados de números que representam cores em cada *pixel*. O algoritmo, portanto, irá associar determinadas sequências numéricas ao rótulo "cão", e outras, ao rótulo "gato". A cada foto do *dataset* consumida, o algoritmo seguirá construindo dois padrões, um para "cão", outro para "gato". Ao final, esse modelo saberá distinguir uma nova foto de cão ou gato que não estava presente em seu *dataset* de treinamento, e, conforme essa foto se aproximar estatisticamente mais de um padrão do que de outro, o modelo indicará que se trata de um "cão" ao invés de um "gato", por exemplo.

Essa aproximação estatística pode ser pensada em termos de distâncias numéricas: suponha que o modelo tenha sido treinado com muitas fotos de cães de pelos pretos e gatos de pelos brancos. Ao apresentarmos a esse modelo pronto uma nova foto de cão de pelagem marrom escura, o agrupamento de *pixels* que representam a cor marrom escuro dos pelos do animal se aproximaria mais do padrão de *pixels* pretos que define um "cão" do que do padrão de *pixels* brancos que define um "gato", simplesmente pelo fato de que o tom de marrom escuro tem uma distância numérica menor em relação ao preto do que em relação ao branco. No entanto, se apresentássemos a esse modelo uma foto de um gato de pelos pretos, qual rótulo ele produziria? É incerta a resposta, dada a natureza numérica do processo. Todavia, espera-se sempre que um modelo discriminativo aplicado tome uma decisão e indique um rótulo, com algum grau de confiança.

16 Tal algoritmo poderia ser uma "rede neural artificial" que utiliza de técnicas eficientes de minimização de erros para buscar o padrão mais comum a todas as fotos de determinado rótulo. Em condições atuais de alto poder computacional, redes neurais com grandes quantidades de variáveis ("neurônios") constituem o chamado *aprendizado profundo* (*deep learning*). Mais especificamente, as redes neurais convolucionais são as mais indicadas para processar imagens, por sua capacidade de abstração hierárquica desses itens.

Imaginemos outro cenário, em que uma máquina dotada de modelo treinado oferece ração para cães e gatos conforme um desses animais se aproxima de sua câmera. Ao conseguir discriminar se aquela imagem é a de um cão ou a de um gato, a máquina decide despejar ração para um ou outro em seu recipiente anexo. Mas o que aconteceria se uma tartaruga se aproximasse dessa máquina? Nesse caso, tal modelo binário não estaria preparado para fazer a devida discriminação; mas, caso a imagem digital da tartaruga tivesse *pixels* que se aproximassem mais dos *pixels* das fotos de cães do *dataset* de treinamento, é possível que essa máquina entregasse ração para cães a uma tartaruga.

Desse simples exemplo, podemos entender que modelos discriminativos, para serem realmente úteis, precisam ser treinados com *datasets* muito maiores e diversos, preferencialmente abrangendo todas as espécies de animais, cada uma com suas diferentes características físicas, em poses e condições de iluminação diversas. Mas é preciso questionar: quem organizaria tamanho *dataset* sem equívocos? Quem garantiria que todas as possíveis representações de todas as espécies de animais seriam contempladas de forma equilibrada? Quem garantiria que o modelo saberia diferenciar, corretamente, cachorros de lobos, dada a semelhança física entre eles? E ainda, se apresentássemos a esse modelo uma imagem de uma árvore, qual seria sua resposta? Ou então, se, em vez de um modelo que tratasse de rótulos concretos (como "cão", "gato", "tartaruga", "lobo"), passássemos a ter um modelo que discriminasse termos mais abstratos (como "criminosos", "suspeitos", "violência", "abuso", "perigo") – como definiríamos esse *dataset*? E, caso fosse treinado, como esse modelo se comportaria? Se concordarmos, todas essas questões nos levam a crer que é impossível existir modelos de IA úteis e perfeitos, dado que haveria muito trabalho humano a ser realizado, propenso a erros e vieses, na constituição desses *datasets*[17].

Há ainda um importante detalhe: os algoritmos de treinamento de modelos discriminativos de imagens costumam utilizar técnicas de "redução de dimensionalidade". Na prática, tais técnicas em reduzir amplamente a quantidade de informações de cada imagem do *dataset* antes de

17 É conhecido o caso da Amazon Mechanical Turk, um serviço da empresa Amazon que reúne pessoas do mundo todo dispostas ao trabalho intenso de rotular dados como textos e imagens em troca de baixíssima remuneração. Consideramos esse tipo de serviço uma precarização do trabalho humano. Ver Crawford e Jolen (2018) e Grohmann (2020).

elas serem processadas para treinamento[18]. Nesses processos, existem perdas de detalhes que contribuiriam para caracterizações mais precisas das imagens. Tudo isso pode nos levar a pensar que os modelos comerciais atuais, ainda que estejam sob o guarda-chuva de grandes empresas de tecnologia internacionais[19] com amplo poder computacional, carregam inúmeros vieses e potenciais problemas em suas capacidades discriminativas. Afinal, são modelos projetados e construídos por (muitos) humanos e subjetividades.

Sugerimos anteriormente que a capacidade discriminativa dos modelos de aprendizado de máquina seria um dos pilares de um suposto processo de criação de máquina: a capacidade de julgamento. Em um primeiro momento, pode parecer pretensioso de nossa parte dizer que uma simples discriminação entre "cão" e "gato" poderia constituir um julgamento em um processo de criação. Mas é justamente essa mesma técnica que pode ser utilizada, também, detecção de padrões em imagens médicas de tomografias, auxiliando na detecção de câncer, por exemplo. Essa mesma técnica pode ser empregada na segmentação de objetos dentro de uma imagem[20]; com isso, é possível fabricar carros autônomos, cujas câmeras conseguem distinguir faixas de uma rodovia, pedestres, calçadas, outros carros, placas de trânsito etc. A mesma técnica pode ser aplicada, ainda, em modelos de reconhecimento facial para detectar padrões únicos de rostos humanos e identificar pessoas em contextos de autenticação e de vigilância. É comum que um modelo discriminativo saiba fazer traduções de um idioma para outro, reconhecer uma música, criar transcrições a partir de um trecho de som ou reconhecer o estilo artístico de desenhos, pinturas e esculturas – tudo é sobre pares de dados e rótulos que, como já sabemos, não precisam ser compostos por apenas uma palavra.

O discriminar de um modelo de IA equivale a ações de classificar, medir, detectar, distinguir e tomar decisões a partir disso. É nesse contexto que ampliamos a noção de julgamento, capacidade atribuída a tais modelos. Seria tal capacidade suficiente para constituir um processo criativo?

18 Segundo o manifesto Nooscópio (Pasquinelli; Joler, 2020), há três camadas de vieses nos processos de aprendizado de máquina: "viés histórico", contido nos próprios dados que representam visões muitas vezes arbitrárias sobre determinados fenômenos ou eventos; "viés de *dataset*", que se insere conforme a interpretação e os potenciais erros de quem o constrói; e "viés de algoritmo", o qual se refere justamente às compressões e perdas de dados das imagens, oriundas dos processos de redução de dimensionalidade, para que o processamento do treinamento dê conta de fazê-lo em tempo aceitável.

19 Aqui é natural citarmos novamente os grandes nomes estadunidenses, como Google, Apple, Amazon, Microsoft, Meta (Facebook), e chineses, como Baidu, ByteDance (TikTok) e Alibaba.

20 Em vez de um rótulo ser uma palavra, ele pode ser um conjunto de coordenadas cartesianas que delimitam o contorno de um objeto.

INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS

Uma vez que um modelo discriminativo é "treinado" por um processo de aprendizado de máquina que consome grandes quantidades de pares de dado (entrada) e rótulo (saída), entendemos que tal modelo se torna capaz de responder, de forma generalizada, qual seria o rótulo de um novo dado, à medida que esse dado se aproxima estatisticamente dos dados de treinamento, já rotulados. O que aconteceria se invertêssemos o processo? Por exemplo, no contexto de modelos que discriminam imagens, em vez de oferecermos uma imagem e esperarmos um rótulo, poderíamos oferecer um rótulo e esperarmos uma imagem como resposta? É precisamente nessa pergunta que contextualizamos a segunda capacidade dos modelos de IA com aprendizado de máquina: a capacidade generativa.

Consideremos que os grandes modelos generativos, como ChatGPT e Dall-E, também são treinados com quantidades massivas de pares de dados e rótulos. Nesse caso, pode nos ser útil a ideia de que um *prompt* seja um conjunto de rótulos, nos quais o modelo irá se basear para gerar algo que se aproxime, estatisticamente, de algumas referências de seu *dataset* de treinamento, relacionadas a tais rótulos. Vamos explorar mais tal percepção.

Dentro do aprendizado de máquina, existem processos de "aprendizado não supervisionado". Em tais processos, considerando determinado conjunto de dados não rotulados, há uma busca automática por padrões numéricos nesses dados e um agrupamento deles, conforme a similaridade entre os padrões encontrados. Tais processos são úteis para encontrar padrões até então não previstos pela supervisão humana[21] e são recorrentes em contextos em que não há a possibilidade de se mapear todas as possíveis categorizações que se poderia atribuir a determinado conjunto de dados. Essa noção é bastante utilizada em sistemas de recomendação: visto que é praticamente impossível categorizar os gostos particulares das pessoas sobre, digamos, filmes e séries, os sistemas de recomendação buscam o tempo todo por novos padrões de comportamento comuns entre as pessoas que assistem a determinado filme ou série. Uma simplificação dessa ideia seria que: levando em conta uma pessoa que assistiu a um filme A e a um filme B, e outra pessoa que também assistiu a A, é possível que o sistema recomende o filme B para essa segunda pessoa citada, por ter detectado um padrão similar entre seus hábitos de consumo e os da primeira pessoa.

21 No aprendizado de máquina, os modelos discriminativos mencionados anteriormente passam por um processo chamado *aprendizado supervisionado*, que trata justamente da constituição de um conjunto de dados que são rotulados por humanos. Tal supervisão se dá no processo de apontar, a uma máquina, o que tal dado é ou não é, para que ela "aprenda" conforme tais indicações.

Obviamente esse exemplo é simplório perto da complexidade que os sistemas de recomendação podem assumir, considerando variáveis como: a quais filmes as pessoas assistiram; qual o gênero dos filmes; suas durações; se as pessoas assistem de uma só vez; se pausam muitas vezes; se abandonam; etc. Tudo isso ainda pode ser cruzado com informações demográficas de região, idade, escolaridade, renda familiar, para se caracterizar, em ainda mais detalhes, os perfis e gostos de usuários.

Para situações complexas como essa, os sistemas de recomendação – modelos baseados em constante aprendizado não supervisionado sobre os dados que capturam de seus usuários – são um dos melhores e mais controversos recursos tecnológicos para se oferecer recomendações precisas, que fazem as pessoas continuarem consumindo.

No contexto dos modelos generativos (como o ChatGPT), quando uma pessoa escreve um *prompt*, digamos, "escreva, em estilo jornalístico, um artigo sobre os conflitos diplomáticos pelos quais o Brasil passou na última década", cada palavra-chave desse pedido se torna um rótulo de algo que, de certo, existe no *dataset* de treinamento do ChatGPT, visto que este é treinado com inúmeros exemplos de artigos e livros publicados na *internet*[22]. A noção de que o modelo não irá resultar em uma referência exata da *internet* (como o buscador Google faz) e irá, em vez disso, gerar um texto inédito, está intimamente relacionada à ideia de se encontrar padrões não previstos entre as referências para os rótulos mapeados em um pedido, de forma não supervisionada.

Podemos compreender que o ChatGPT habilidosamente buscará, em seu imenso espaço de representação digital (numérica), aproximações com textos relacionados a "estilo jornalístico", "artigo", "Brasil", "diplomacia", "conflitos", "última década". O processo de aprendizado não supervisionado poderá encontrar novos padrões entre os textos que foram rotulados de forma supervisionada conforme tais palavras-chave, chegando a um espaço de representação combinatório, o qual o modelo generativo poderá explorar para alcançar um resultado plausível e aparentemente inédito. Mais especificamente, modelos generativos como ChatGPT constroem textos palavra por palavra; a cada nova palavra, os modelos exploram possíveis próximas palavras dentro de um determinado espaço de representação, restrito a partir dos dados relacionados às palavras-chave extraídas do *prompt*.

22 Segundo Brown *et al.* (2020, p. 8), um *dataset* para modelos como ChatGPT possui 45TB de texto puro, não filtrado, oriundo de textos de páginas na *internet* e livros. Essa massa de dados equivale a 45 trilhões de caracteres, o que corresponderia a, aproximadamente, 90 milhões de livros diferentes (assumindo que cada livro possui em média cem mil palavras e que cada palavra possui em média cinco caracteres).

Um entendimento importante aqui é que a geração se baseia em aproximações com os dados de treinamento e, como já sabemos, é praticamente impossível construir um *dataset* perfeito, sem vieses. Nesse sentido, quando esses modelos exploram espaços de representação que não possuem referências bem definidas ou diversificadas em relação ao que se busca, informações inverossímeis podem ser produzidas[23].

Como outro exemplo, podemos pensar nas maneiras pelas quais modelos como Stable Diffusion, Midjourney ou Dall-E chegam a imagens inéditas e plausíveis por meio desse mesmo recurso. Já sabemos que uma imagem é constituída de uma sequência de números que representam cores em *pixels*. Para cada imagem, ou "massa de *pixels*", atribuímos rótulos, dentro de um processo de aprendizado supervisionado. Quando uma pessoa escreve um *prompt* pedindo, digamos, "crie uma cena de um cartum popular dos anos 1990 cujos personagens possuem pele amarela", um modelo generativo de imagens também irá relacionar palavras-chave ("cartum", "anos 1990", "pele amarela") com espaços restritos de representação de imagens, obtidos por aprendizado não supervisionado.

Há um termo mais interessante para esses espaços restritos de representação das imagens: espaços latentes. Um espaço latente compreende um espaço de representação codificado e comprimido[24] entre pares de imagens e rótulos, que capturam correlações entre textos e padrões visuais. O modelo explora tais espaços latentes, formados por aprendizado não supervisionado sobre o conjunto de palavras-chave do *prompt* e respectivas imagens, chegando a resultados que combinam padrões de *pixels* (como em uma colagem), mas de forma plausível, para nos convencer de que a imagem gerada tem a integridade de uma fotografia, de um desenho ou de uma ilustração.

Os exemplos de *prompt* sobre texto jornalístico acerca da diplomacia brasileira no ChatGPT e sobre cartum dos anos 1990 nos geradores de imagem foram propositalmente escolhidos: se retomarmos os processos judiciais movidos contra as empresas mantenedoras desses modelos, lembraremos que ambos indicam violações de direitos autorais. O *The New York Times* indica que o ChatGPT viola seus direitos por utilizar vários dos artigos do famoso veículo de imprensa em seu *dataset* de treinamento.

23 Não à toa, o ChatGPT sempre exibe a seguinte mensagem de isenção de responsabilidade no rodapé de sua interface: "ChatGPT can make mistakes. Consider checking important information" ("O ChatGPT pode cometer erros. Considere checar informações importantes").

24 Podemos aplicar a noção de *redução de dimensionalidade* aqui também, uma vez que uma imagem de milhões de *pixels* e seu respectivo rótulo podem ser reduzidos a códigos mais simples e de menor tamanho.

Já a classe de artistas visuais se queixa do Stable Diffusion, Midjourney e Dall-E por utilizarem, em seus *datasets,* imagens autorais e por chegarem a resultados que muito se assemelham a tais imagens. Se observarmos os exemplos trazidos por Marcus e Southen (2024), os experimentos com *prompts*, bastante semelhantes aos *prompts* aqui exemplificados, geraram artigos que seriam claramente plágios de artigos escritos por jornalistas do *The New York Times*, além de imagens que poderiam ser consideradas cópias fiéis de cenas do famoso cartum dos anos 1990 *Os Simpsons*, de Matt Groening. Por que um modelo generativo comercial chegaria a essa condição de plágio se supostamente gera produtos inéditos, que inclusive dão direitos comerciais aos assinantes do serviço?

Como já discutimos, um *dataset* de treinamento que não represente uma boa diversidade de artigos jornalísticos sobre diplomacia brasileira, ou de imagens de cartuns dos anos 1990 com personagens de pele amarela, tende a ser enviesado. Em espaços latentes formados com poucas referências, ou com um *prompt* muito específico, a tendência é que tais modelos gerem produtos muito semelhantes aos textos e às imagens constantes em seu *dataset* de treinamento. Isso pode fazer com que qualquer pessoa que não conheça as obras originais utilize e comercialize inadvertidamente tais produtos generativos, numa cadeia de violação de direitos autorais. Tais violações ocorrem não apenas porque as empresas lucram sobre a utilização de imagens autorais sem devida autorização em *datasets,* mas também porque os modelos podem produzir plágio.

Consideremos, portanto, que tanto a característica discriminativa quanto a generativa dos grandes modelos contemporâneos de IA são controversas em um sentido ético. Elas reproduzem senso comum, vieses e estereótipos: se digitarmos apenas "médico" em um *prompt* de imagem, há grandes chances de recebermos a imagem de um homem caucasiano com jaleco branco, o que tende a ignorar a existência de médicos de outras raças e gêneros. Se formos muito específicos em uma descrição de pedido, com palavras-chave mal representadas no *dataset*, podemos incorrer em casos de plágio. Em um extremo ou outro, os modelos podem gerar informações inverossímeis, potencializando notícias falsas, através de textos e imagens com alto poder de convencimento.

A capacidade generativa, segundo pilar de um suposto processo criativo de máquina, populariza-se hoje, conforme já abordado, por intermédio de modelos como o ChatGPT e o Stable Diffusion. Podemos compreender que, para qualquer geração de máquina, há também um julgamento de máquina. O processo de aprendizado é justamente constituído desse ciclo: ao longo do tempo, um modelo em formação pode gerar imagens e textos, que, por sua vez, são comparados às referências do *dataset* de treinamento.

Enquanto um modelo julgador for capaz de discriminar que determinada geração não condiz com um resultado satisfatório, o modelo generativo deve se aperfeiçoar até que consiga atender aos critérios e às métricas do modelo julgador. Tal aperfeiçoamento se dá de forma matemática, através de funções de erro que comparam dados gerados a dados presentes em *dataset* de treinamento. O modelo melhora à medida que ficarem menores as diferenças entre o gerado e a referência, a cada ciclo de ajustes e novas comparações. Resta-nos saber se esse tipo de processo automatizado nos permite atribuir agência e criatividade às máquinas e, ademais, quanto isso pode ser útil para artistas em suas atividades criativas.

ARTISTAS E INTELIGÊNCIAS ARTIFICIAIS: COCRIAÇÃO?

Margaret Boden (2011) nos explica que a ação criativa, dentro do espaço conceitual e cultural de um indivíduo, pode assumir três diferentes vertentes: combinatória, exploratória e transformadora.

A forma combinatória implica em "combinações não familiares de ideias familiares" (Boden, 2011, p. 31) para a geração de novas possibilidades entre coisas que já permeiam nossos cotidianos e repertórios criativos. A forma exploratória, por sua vez, corresponderia a buscar regiões "desconhecidas" dentro de um espaço conceitual, o que implica muitas vezes em observar ideias conhecidas, explorando-as e as experimentando sob outras perspectivas. Já a forma transformadora envolve modificar o próprio espaço conceitual, expandindo-o, criando relações e analogias, fora da conformidade de um espaço conceitual dado até determinado momento.

Sob essa teoria de Boden e considerando o que aprendemos sobre os modelos discriminativos e generativos, no máximo poderíamos dizer que tais modelos são criativos em forma combinatória. Pares de dados e rótulos, exploração de espaços latentes ou ciclos de geração e julgamento não fazem com que esses modelos alcancem outras perspectivas sobre os conceitos, para além daquelas predeterminadas por um *dataset* de treinamento. Tampouco vemos tais modelos expandindo espontaneamente seus espaços conceituais, pois são limitados a trazer apenas referências entre aquelas definidas pelos seus *datasets*.

E, se expandirmos a noção de criatividade para algo que envolve intencionalidades, não poderíamos apenas adaptar a teoria de Boden para modelos generativos e dizer que são criativos simplesmente porque são eficientes em combinar. Em pesquisas passadas, exploramos a ideia de que, para ser considerada verdadeiramente criativa, uma máquina deveria conseguir ter autonomia para modificar espontaneamente suas significações

e seus objetivos, baseando-se inclusive em noções de lazer, prazer e autorrealização (Venancio Júnior, 2019a, p. 198). Claramente, os modelos generativos mais atuais não são projetados para refletir tais noções, já que os próprios interesses comerciais das empresas não as contemplam. Os modelos operam em ciclos de geração e julgamento bastante condicionados à correlação e combinatória de padrões constantes em seu *dataset* de treinamento, que sabemos ser geralmente limitado por vieses e senso comum. Se entendemos tais processos, qualquer impressão de atribuição de significados ou quaisquer mudanças de objetivos que o modelo de IA possa causar não passariam de coincidente exploração de espaços latentes entre conteúdos convincentes acerca de tais ideias.

Isso remonta ao famoso Teste de Turing, ou o Jogo da Imitação (Turing, 1950), em que uma máquina poderia ser considerada inteligente se, em uma conversa por trocas de mensagens de texto, conseguisse convencer um humano de que ela não seria uma máquina, mas sim outro humano. Analogamente, poderíamos nos convencer das fantásticas narrativas que são geradas em torno de modelos como ChatGPT e afins – de máquinas com vontade própria, que se declaram inteligentes ou criativas, ou mesmo que estão tornando humanos obsoletos em atividades criativas – somente se ignorarmos os bastidores desses sistemas, os funcionamentos de seus algoritmos e os conjuntos de dados que os formam.

Questionamos, portanto, a noção de *cocriação* nesse contexto, partindo da ideia de que cocriar seria equivalente a reunir mais de uma entidade criativa para um trabalho conjunto. Se uma máquina não é criativa na extensão aqui discutida, podemos afirmar que, no máximo, artistas dispõem de uma ferramenta impressionante em suas capacidades de geração de combinações de coisas prováveis e improváveis[25]. Mas é válido lembrar que, quando trabalhamos com modelos treinados (como Dall-E, Stable Diffusion e Midjourney), restringimo-nos a combinações de produções de senso comum, baseadas na representatividade que imagens populares e estereótipos possuem em *datasets* robustos, como o LAION-5B. É esse o lugar comum de exploração criativa de que um artista dispõe ao trabalhar com tais modelos generativos treinados. Seria-nos muito caro chamar tal processo de *cocriação*. Poderíamos até lembrar que nenhuma tecnologia é fruto senão de uma intenção e de um trabalho humano. Se considerarmos os modelos discriminativos e generativos comerciais como resultados de intenções humanas, restaria que nos perguntássemos se tais intenções poderiam ser agregadas a um processo de cocriação – o que nos parece

25 Como exemplo, a própria OpenAI propagandeou seu modelo Dall-E com imagens de "sofás em formato de abacate" ou "astronautas cavalgando unicórnios no espaço sideral". Ver: https://openai.com/dall-e-2. Acesso em: 13 abr. 2024.

muito restrito em termos de coerência entre estratégias criativas, dado que um artista teria intenções muito subjetivas, que dificilmente dialogariam na medida certa com as intenções comerciais que estão por trás do emprego de algoritmos de aprendizado e da organização de *datasets* para amplo alcance de público.

Nossa reflexão, neste momento, volta-se para questionamentos sobre como artistas poderiam de fato utilizar tais tecnologias de IA a favor de sua própria criatividade. Oferecemos a ideia de que o artista tenha maior controle sobre o processo de treinamento de modelos de IA, manipulando *datasets* e algoritmos de aprendizado, para chegar a modelos que se comportem em prol de suas próprias intencionalidades.

Trazemos o exemplo da artista sino-canadense Sougwen Chung, que desenvolve robôs para capturar e reproduzir seus gestos em desenhos e pinturas[26]. A artista performa com seus robôs treinados, que decerto discriminam contextos e situações dentro da performance para gerarem, por si só, novos gestos, geralmente amparados por instrumentos de desenho ou pintura. A ação conjunta entre Chung e seus robôs nos projetos *Drawing Operations* (2015-2018) resulta em desenhos e pinturas que refletem um processo de retroalimentação entre movimentos espontâneos da artista e movimentos quase imprevisíveis das máquinas, de modo que Chung responde aos traços dos robôs, e os robôs, aos traços dela, formando composições inéditas. Chung chegou a declarar que suas composições são "colaborações", conjugando modos de ver e traçar distintos entre humanos e máquinas, entre o orgânico e o sintético[27].

Os exemplos de Chung podem nos trazer um norte de possibilidades, considerando sua investida em controlar as IAs à medida que explora as características intrínsecas dessas tecnologias, desde potencialidades até limitações dos modelos e processos de aprendizado, incluindo restrições físicas dos robôs. E quando a artista se dispõe a colaborar com máquinas, seu processo criativo passa a sofrer interferências de tais tecnologias, o que nitidamente é intencional. Chung busca novas experiências e resultados a partir do momento em que estabelece tais diálogos. Isso não implica em demandar que uma máquina seja criativa nesse processo, mas sim em explorar profundamente suas possibilidades de julgar e gerar, em um contexto mais limitado e controlado.

26 Ver: https://sougwen.com/. Acesso em: 13 abr. 2024.
27 Ver: https://sougwen.com/machinecollaboration. Acesso em: 13 abr. 2024.

CONSIDERAÇÕES FINAIS

Buscamos com esta reflexão compreender melhor os motivos dos movimentos litigiosos contra as empresas mantenedoras do ChatGPT, do Co-Pilot e do Stable Diffusion. Se tais modelos dependem de *datasets* que consomem grandes quantidades de produções vinculadas a direitos autorais, mas são comercializados sob a premissa de serem criativos e gerarem produções inéditas, entendemos que isso é improcedente de diversas maneiras.

A natureza dos processos de associação pareada entre dados e rótulos é a chave para compreendermos como se dá um processo de julgamento de máquina, em que se analisa um padrão de textos ou imagens para se buscar o rótulo associado aos padrões que mais se aproximam dele. O processo inverso, de geração de padrões a partir de rótulos, ocorre pela exploração de espaços latentes. Tais espaços são enviesados pelos aspectos culturais em que se inserem, pela seletividade do *dataset* e pela própria característica algorítmica de extração de padrões, e não de diferenças. A utilização de vastos conjuntos de dados, que incluem obras protegidas por direitos autorais, para treinar esses modelos generativos, destaca um desafio crítico na regulação dessas tecnologias, que ficam sujeitas a protestos.

As ações legais sublinham a necessidade de uma clara legislação que equilibre inovação e proteção criativa, garantindo que os avanços das IAs não prejudiquem os direitos dos criadores originais. Além disso, esses casos estimulam um diálogo necessário sobre a verdadeira natureza da criatividade das máquinas e o papel do humano na era da criação digital, desafiando-nos a refletir sobre como as tecnologias de IA devem evoluir e ser regulamentadas em um mundo cada vez mais digital e interconectado.

REFERÊNCIAS

BODEN, Margaret A. *Creativity and Art.* Three Roads to Surprise. Oxford: Oxford University Press, 2011.

BROWN, Tom B. *et al*. Language Models are Few-Shot Learners. *In: arXiv*, 22 jul. 2020. Disponível em: https://arxiv.org/pdf/2005.14165.pdf. Acesso em: 13 abr. 2024.

BUTTERICK, Matthew. GitHub Copilot Litigation, 2022. Disponível em: https://githubcopilotlitigation.com/. Acesso em: 5 jan. 2024.

BUTTERICK, Matthew. Stable diffusion litigation, 2023. Disponível em: https://stablediffusionlitigation.com/. Acesso em: 5 jan. 2024.

CRAWFORD, Kate; JOLEN, Vladan. Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources, 2018. Disponível em: https://anatomyof.ai. Acesso em: 13 abr. 2024.

DESCARTES, René. *Discurso do método*. Tradução de Paulo Neves. Porto Alegre: LP&M, 2017.

GROHMANN, Rafael. Plataformização do trabalho: entre a dataficação, a financeirização e a racionalidade neoliberal. *Revista Eptic*, v. 22, n. 1, 2020. Disponível em: https://seer.ufs.br/index.php/eptic/article/view/12188. Acesso em: 13 abr. 2024.

GRYNBAUM, Michael M.; MAC, Ryan. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*, 27 dec. 2023. Disponível em: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. Acesso em: 4 jan. 2024.

MARCUS, Gary; SOUTHEN, Reid. Generative AI Has a Visual Plagiarism Problem – Experiments with Midjourney and DALL-E 3 show a copyright minefield. *IEEE Spectrum*, 6 jan. 2024. Disponível em: https://spectrum.ieee.org/midjourney-copyright. Acesso em: 13 abr. 2024.

PASQUINELLI, Matteo; JOLER, Vladan. O manifesto Nooscópio: inteligência artificial como instrumento de extrativismo do conhecimento. Tradução de Leandro Módolo e Thais Pimentel. *Lavits*, 1 maio 2020. Disponível em: https://lavits.org/o-manifesto-nooscopio-inteligencia-artificial-como-instrumento-de-extrativismo-do-conhecimento/. Acesso em: 13 abr. 2024.

TURING, Alan M. Computing Machinery and Intelligence. *Mind*, New Series, v. 59, n. 236, p. 433-460, 1950.

VENANCIO JÚNIOR, Sergio J. Arte e inteligências artificiais: implicações para a criatividade. *ARS (São Paulo)*, [s. l.], v. 17, n. 35, p. 183-201, 2019. Disponível em: https://www.revistas.usp.br/ars/article/view/152262. Acesso em: 14 out. 2023.

VENANCIO JÚNIOR, Sergio J. Extentio: desenhos de máquina, desígnios humanos. 2019. 200 p. Dissertação (Mestrado em Artes Visuais) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2019.