Inteligência artificial e enunciação: análise de grandes coleções de imagens e geração automática via Midjourney

Maria Giulia Dondero*
https://orcid.org/0000-0003-2320-8130

Tradução, revisão e notas:
Gustavo H. R. de Castro**
https://orcid.org/0000-0003-4486-9579

Matheus Nogueira Schwartzmann***
https://orcid.org/0000-0002-2887-3570

Como citar este artigo: DONDERO, M. G. Inteligência artificial e enunciação: análise de grandes coleções de imagens e geração automática via Midjourney. Tradução, revisão e notas Gustavo H. R. de Castro e Matheus Nogueira Schwartzmann. *Todas as Letras – Revista de Língua e Literatura*, São Paulo, v. 26, n. 2, p. 1-24, maio/ago. 2024. DOI: https://doi.org/10.5935/1980-6914/eLETTR17164.

Submissão: 10 de junho de 2024. **Aceite:** 4 de julho de 2024.

Resumo: As inteligências artificiais (IA) têm simulado, cada vez mais satisfatoriamente, a particularidade da linguagem e das ações humanas. É, portanto, imprescindível que a semiótica trate dessas tecnologias e práticas de automatização. Com o intuito de realizar essa abordagem, assumiremos o ponto de vista da teoria da enunciação de Émile Benveniste, especialmente de três desenvolvimentos em semiótica aplicados ao estudo da IA, que nos permitirão discutir: 1. a tradução entre linguagem artificial e linguagem humana; 2. a relação entre

* Université de Liège (ULiége), Liège, Bélgica. *E-mail*: mariagiulia.dondero@uliege.be

** Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), Araraquara, SP, Brasil. Fapesp n. 19/27000-7, Capes-PrInt. *E-mail*: g.castro@unesp.br

*** Unesp, Araraquara, SP, Brasil. Capes-PrInt. *E-mail*: matheus.schwartzmann@unesp.br

banco de dados, algoritmos e criatividade maquínica, por meio da análise de grandes coleções de imagens (*big visual data*); e 3. os modos de endereçamento da máquina ao usuário do modelo de inteligência artificial generativa Midjourney.

Palavras-chave: *Big visual data*. Enunciação. Geração automática de imagens. Inteligência artificial. Midjourney.

Considerações iniciais[1]

A história da inteligência artificial (IA) remonta à década de 1950 e à máquina de Turing[2], que continua sendo o modelo teórico fundamental de toda computação atual: esse foi o início da digitalização e da automatização dos cálculos. Se dermos, ainda, um salto no tempo rumo à história mais recente, veremos que a automatização da computação está na base do funcionamento de diversos tipos de utilitários cotidianos: mecanismos de busca, sistemas de recomendação de produtos e de navegação, jogos de estratégia, *chatbots* e, mais recentemente, os modelos de geração automática de imagens[3], que são o tema deste artigo.

Desde então, o termo IA tem sido aplicado a diferentes tecnologias, que foram superadas e/ou aperfeiçoadas, desmembrando-se em ramificações cada vez mais complexas. Nesse cenário, erigiram-se duas grandes abordagens: uma *simbólica* e outra *conexionista*. A primeira dominou a pesquisa e o fomento desde o final de 1950 até meados de 1990. Ancorada nas tradições da lógica matemática, da engenharia de sistemas e da cibernética, a abordagem simbólica partia da ideia de que os computadores podiam reproduzir certos aspectos racionais do pensamento humano (resolver problemas, fazer julgamentos, tomar decisões) por meio de programas de processamento de símbolos (baseados em regras e combinações). A segunda abordagem, por sua vez, é aquela que predomina atualmente nas pesquisas – e a que nos interessa aqui. Ancorada em uma tradição que integra matemática, estatística, psicologia cognitiva e neurociências, a abordagem conexionista se baseia na ideia de que as IAs podem ser concebidas como uma forma de *aprendizagem automática*, realizada a partir de algoritmos computacionais capazes de aprender a reconhecer padrões em grandes arquivos, por meio de sequências indutivas de tentativas e erros, a fim de fazer previsões sobre dados recém-fornecidos, imitando, de maneira simplificada e esquemática, as conexões entre os neurônios: daí a origem do nome "redes neurais artificiais".

1 Agradecemos a Adrien Deliège por suas explicações sobre métodos de análise em aprendizado de máquina e a Marion Colas-Blaise e Jean-François Bordron, pela releitura deste texto. Também agradecemos a Enzo D'Armenio por seus comentários sobre a geração de imagens usando o Midjourney e a Jean Cristtus Portela e a Renata Mancini, pelos convites que nos permitiram discutir essas ideias na Faculdade de Ciências e Letras, *c*ampus de Araraquara, da Universidade Estadual Paulista "Júlio de Mesquita Filho" (FCLAr-Unesp) e na Universidade de São Paulo (USP), respectivamente.

2 A máquina de Turing consiste na metáfora conceitual de uma fita infinita, que atua como memória de longo prazo, na qual símbolos podem ser lidos e escritos, e de uma cabeça de leitura/escrita que se move ao longo da fita, segundo uma tabela de instruções responsáveis por determinar as operações (N. T.).

3 As primeiras tentativas de gerar imagens automaticamente datam dos anos 1960-1970, com o programa Aaron, de Harold Cohen. Posteriormente, uma série de tecnologias foram desenvolvidas. Apenas a título de exemplo, podemos mencionar alguns marcos: em 2018, surgiram as GANs progressivas, seguidas pelo *BigGAN*, da Google, que permitiram gerar imagens aprimorando-as gradualmente em termos de resolução. Em 2021, a OpenAI introduziu o DALL-E, inaugurando a geração de imagens a partir de descrições textuais (N. T.).

A despeito dessas abordagens e dessa diversidade de tecnologias, o que tem caracterizado as IAs – e o que nos interessa enquanto semioticistas – é o fato de elas oferecerem ferramentas que tentam simular, de modo cada vez mais convincente e capilar, a particularidade da linguagem do ser humano e de suas práticas, inclusive as práticas de pensamento[4]. Por isso, é imprescindível que a semiótica pós-estruturalista trate das linguagens artificiais e das tecnologias e práticas de automatização das ações humanas.

Segundo o ponto de vista da semiótica, portanto, podemos estabelecer ao menos três abordagens da enunciação que podem ser destinadas ao estudo da IA: 1. enunciação como endereçamento; 2. enunciação como mediação entre linguagem maquínica e linguagem natural; e 3. enunciação como práxis enunciativa. Principalmente no caso da segunda abordagem, se, tradicionalmente, essa teoria tem lidado com a mediação entre a linguagem – entendida como um sistema de virtualidades/possibilidades – e o discurso – entendido como a realização dessas virtualidades/possibilidades –, a conversão entre a linguagem maquínica e a linguagem humana só pode dizer respeito ao ato de desvinculação entre algo possível e manipulável e os produtos da linguagem[5] ou, ainda, no caso das IAs ditas generativas[6], ao ato de criar produtos linguageiros a partir de bancos de dados.

Essas três perspectivas sobre a enunciação nos permitem: 1. compreender o modo como a máquina se dirige ao usuário do Midjourney e do ChatGPT e, principalmente, aos robôs; 2. descrever a tradução entre linguagem artificial e linguagem humana; e, por fim, 3. entender a relação entre banco de dados, funcionamento de algoritmos e criatividade maquínica, por meio do *estudo da relação entre inovação e sedimentação* em grandes coleções de imagens, consideradas como arquivos e como locais onde o novo é gerado.

Para começar, na primeira parte deste artigo, trataremos da segunda abordagem da teoria da enunciação, a saber, a mediação entre linguagem maquínica e linguagem humana. Na sequência (primeira seção da segunda parte deste artigo), abordaremos a terceira concepção de enunciação (práxis enunciativa) buscando, especialmente, estudar a relação entre bancos de dados e trabalho algorítmico. Nesse último caso, em especial, discutiremos, de início, a *análise* de bancos de dados[7] a fim de examinar o modo como a visão computacional (*computer vision*), conjuntamente com outras disciplinas como a história da arte, permite a *análise de grandes quantidades de dados* (os *big visual data*) usando, para isso, algoritmos apropriados que transformam análises estatísticas em visualizações de imagens (metaimagens). Em seguida (segunda seção da segunda parte deste artigo), nos dedicaremos ao estudo da *geração automática de enunciados* visuais, isto é, as grandes coleções arquivadas em bancos de dados, usadas

4 Na década de 1950, Turing fez uma pergunta fundamental em seu famoso artigo "Computing machinery and intelligence" (1950): "A máquina pode pensar?". Para uma discussão filosófica sobre a origem, a história e os desenvolvimentos da máquina de Turing, ver Lassègue (2017).

5 Sobre algoritmos como um actante *coletivo*, ver Dondero (2019a).

6 As IAs ditas generativas (ou gerativas) são sistemas que criam conteúdos autonomamente, como texto, imagens, música e códigos, a partir da aprendizagem de padrões identificados em grandes conjuntos de dados. Esses sistemas são chamados de generativos porque criam informações em vez de apenas analisarem dados (N. T.).

7 A análise de bancos de dados, em especial de grandes coleções de imagens (*big data*), é o tema central do nosso projeto de pesquisa atual, iniciado em 2022 e intitulado *Towards a genealogy of visual forms: semiotic and computer-assisted approaches to large image collections* (FRS-FNRS-Bélgica). Mais informações estão disponíveis em: https://ceserh.hypotheses.org/p-d-r-towards-a-genealogy-of-visual-forms.

para produzir novos enunciados a partir de textos antigos, já sedimentados na memória coletiva, por meio de operações ou mesmo de instruções (*prompts*)[8]. Especialmente no caso da geração de novos enunciados, estudaremos algumas interações e alguns de seus produtos textuais obtidos, sobretudo, por meio do Midjourney.

Tradução entre linguagem artificial e linguagem natural

Consideremos esta pergunta, por sinal, bastante simples: "Por que a semiótica deve abordar a inteligência artificial?". A primeira resposta, a mais banal, mas a mais fundamental, poderia ser: a IA coloca o problema da tradução entre linguagens no centro de seus experimentos. Não falamos, aqui, apenas da tradução entre linguagem verbal e linguagem visual – questão que abordaremos mais adiante, quando analisarmos as produções do Midjourney. Falamos, especialmente, de um primeiro tipo de tradução, bem mais fundamental, que ocorre entre o referente maquínico (binário, composto por zeros e uns) e a linguagem natural, por meio de uma linguagem artificial.

De modo geral, partimos de uma definição simples do que seria uma linguagem artificial: um sistema de expressão construído a partir de signos linguísticos e/ou outros símbolos, visando reduzir a ambiguidade e a variabilidade inerentes às linguagens naturais, e regido por regras de uso explícitas e prescritivas. No campo da computação, essas linguagens artificiais, cujas regras são explicitamente estabelecidas antes do uso e baseadas na matemática, são chamadas de "linguagens de programação". Transpondo a problemática para as ciências da linguagem e, portanto, para o aspecto propriamente humano da questão, devemos começar pelo estudo da tradução entre linguagem maquínica e linguagem natural, em termos de enunciação e de conversões entre os níveis do percurso gerativo (Greimas, 1983), do mais abstrato para o mais "figurativo" e vice-versa. Especialmente no nosso caso, nos referimos aos seguintes níveis: 1. da máquina (referente); 2. da mediação das linguagens de montagem; 3. das linguagens de programação; e 4. e da linguagem natural. Passemos, então, a essa tarefa.

Enunciação como mediação em IA

No que concerne à enunciação como *mediação*, seguiremos o artigo de Andrea Valle e Alessandro Mazzei (2017), que identifica três níveis de pertinência semiótica, estratificados de acordo com seu grau de abstração e de figuratividade.

8 No contexto da computação, um *prompt* consiste em instruções ou estímulos dados, por exemplo, a sistemas de IA para gerar respostas ou realizar tarefas específicas, direcionando o modelo na produção de conteúdo (N. T.).

DOI 10.5935/1980-6914/eLETTR17164

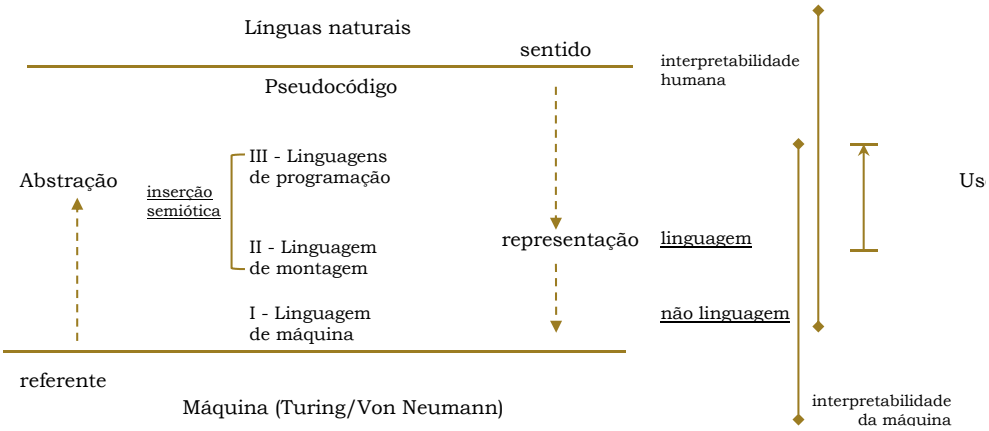

Figura 1 – Diagrama de Valle e Mazzei

Fonte: Valle e Mazzei (2017, p. 509, tradução nossa).

Esse diagrama demonstra, justamente, a estratificação em três linguagens. A *linguagem de máquina* assume a forma de um sistema binário de signos, uma alternância/combinação de zeros e uns. Trata-se, segundo Tanaka-Ishii (2010, p. 25, tradução e grifos nossos), "[d]o único sistema de sinais de larga escala que possui um intérprete totalmente *explícito* e totalmente *caracterizado*, externo ao sistema interpretativo humano". A *linguagem de montagem*[9], responsável pela mediação entre a linguagem de máquina e a linguagem de programação, que já é constituída por um primeiro recobrimento linguístico inicial (o *envelope linguístico*). E, por fim, a própria *linguagem de programação*, isto é, o código ou a forma de escritura em questão.

De acordo com a definição de Turing (1936, p. 230-265), uma linguagem de programação é uma linguagem formal implementada em uma máquina abstrata, que executa operações de leitura e de escrita em uma memória infinita. A partir dessa definição, já podemos ter uma ideia da relação entre as três linguagens mencionadas e suas conversões[10]. Ora, é evidente que o referente da máquina é apenas um sistema baseado na lógica matemática. Trata-se de um sistema semiótico, totalmente inequívoco, mas não de um sistema linguístico. Para Valle e Mazzei (2017), estamos diante do sistema de linguagem entendido como referente abstrato[11].

Um olhar mais atento logo revela interpretações mútuas nesse jogo entre a máquina – enquanto referente, local de armazenamento para uma memória infinita e a linguagem natural. Isso, é claro, se pudermos entender a interpretação como uma "cadeia de substituições" ou como uma conversão entre níveis.

9 Linguagens de montagem (*assembly languages*) são, grosso modo, linguagens de baixo nível (baixo, pois estão mais próximas do *hardware* e do código de máquina do que das abstrações oferecidas pelas linguagens de alto nível), que fornecem uma interface direta entre o código escrito pelo programador e o conjunto de instruções do *hardware*, daí o seu papel mediador (N. T.).

10 Sobre essa *estratificação*, que é gerada pela possibilidade de decompor toda a escrita em uma *multiplicidade de níveis que se comunicam entre si*, desde a linguagem de máquina até níveis mais lógicos, ver Jean Lassègue (2017), especialmente quando o autor discute essa estratificação e sua relação com a escrita alfabética, que é, por sua vez, dotada de "uma única camada".

11 Em uma entrevista de Andrea Valle concedida a Basso e Dondero (2019a, tradução nossa), Valle aponta que a máquina, enquanto um dispositivo de escrita e de leitura, é capaz de converter uma representação em uma prática. O problema não é a representação em si, mas sim o comportamento resultante.

Vejamos isso mais detalhadamente: do ponto de vista humano, do programador, estamos diante de uma conversão do nível linguístico, de superfície, para variações do estado da memória física (nível mais profundo), que garantem uma interpretação inequívoca. Disso decorre que o primeiro nível do diagrama, *o da máquina* (o referente, ver Figura 1), deve, necessariamente, consistir em uma memória e em um dispositivo capaz de manipulá-la, composto, por sua vez, de um conjunto de locais discretos de memória, dependentes do *design* do processador. É verdade que, na máquina, as operações se realizam por meio da codificação executada pelo processador, que lê instruções compostas por uma palavra binária indicando a operação, e por um número, oriundo de outro código binário, responsável por determinar um endereço dentre todos os endereços de locais de memória possíveis. Mas é a *linguagem de montagem* que introduz, de fato, uma forma de representação linguística para identificar as operações envolvidas ("armazenar", "carregar", "adicionar" etc.), por meio de um sistema de nomes associados a grupos de instruções ou endereços de memória. Nesse segundo nível do diagrama, são controladas as possibilidades de operações em locais de memória. Por fim, as *linguagens de programação* estão posicionadas no terceiro nível do diagrama, pois são responsáveis por inserir instruções de controle de fluxo e por implementar construções de uma linguagem específica, que podem diferir significativamente de um "dialeto" de programação para outro. De fato, como afirma Andrea Valle (2019b) – em artigo dedicado a um dos primeiros programas de computador, o *Beginner's All-Purpose Symbolic Instruction Code* (Basic)[12] –, há uma proliferação de dialetos a partir da versão básica desse código (o Basic), que fornecem ordens de execução (*statements* ou instruções) parcialmente diferentes e sintaxes mutuamente incompatíveis.

Uma "dialetização" desse tipo significa que os programas de dialeto só podem ser executados dentro da estrutura da comunidade de programadores do dialeto em questão. Nesse sentido, como Valle (2019b) aponta, há uma espécie de *vade mecum*, o *The Basic handbook* (Lien, 1981), que visa possibilitar a conversão de um dialeto para outro entre os milhares disponíveis. Além disso, ele também fornece uma vasta literatura sobre os estilos de programação em geral. Por exemplo, *Elements of programming style* (Kernighan; Plauger, 1978) com seus verdadeiros conselhos de estilo ou *Exercises in programming styles* (Lopes, 2014), obra que revisa a noção de estilo retirada de *Exercises in style* (Queneau, 1947), para discutir os estilos restritos[13] na ciência da computação[14], chamando a atenção para a importância de os programas serem bem escritos não tanto para a máquina, mas sobretudo para o usuário[15]. De fato, Valle (2019b) também argumenta que as três linguagens de programação que ele considera em sua reflexão – quais sejam: Python, SuperCollider e Basic – estão associadas ao mesmo plano de conteúdo no nível da linguagem maquínica. Obviamente, "mesmo" deve ser entendido do ponto de vista da interpretação da máquina, que é com-

12 Esse código está em operação desde 1964, graças a John G. Kemeny e Thomas E. Kurtz, e se tornou a linguagem de programação padrão para computadores na década de 1970. A esse respeito, ver Kemeny e Kurtz (1964).

13 Os estilos restritos referem-se a abordagens, práticas ou padrões destinados a limitar a maneira como o código é escrito – ou como os sistemas são projetados. No caso de Lopes (2014), a ideia é explorar como diferentes estilos podem impactar a clareza, a manutenibilidade e a eficiência do *software* (N. T.).

14 Segundo Lopes (2014, p. xxii, tradução e grifos nossos): "podemos escrever uma variedade de programas que são virtualmente idênticos em termos do *que fazem*, mas que são radicalmente diferentes em termos de *como o fazem*".

15 Outro texto importante sobre a codificação como uma forma de "escrita literária" é a obra de Knuth (1984).

DOI 10.5935/1980-6914/eLETTR17164

pletamente inequívoca: todas essas linguagens podem ser consideradas equivalentes do ponto de vista da máquina, mas não do ponto de vista dos usuários.

Nesse cenário, o objetivo da semiótica deve ser estudar não apenas as *conversões* entre níveis, isto é, entre a linguagem enquanto referente abstrato da máquina e a linguagem natural, mas igualmente os diferentes estilos retóricos de cada linguagem de programação[16]. Valle e Mazzei (2017) também argumentam que, pelo fato de todas essas linguagens de programação serem capazes de expressar a "mesma coisa" (os mesmos estados e as mesmas operações em relação à máquina, entendida como um sistema de referência), a sua *variedade* se explica, justamente, pela necessidade de expressar melhor essas "mesmas coisas", isto é, de modo mais fácil, eficiente, elegante, claro, rápido: em suma, algumas coisas são mais bem expressas por algumas linguagens de programação do que por outras[17].

Contudo, se quisermos falar do ponto de vista do usuário, é preciso considerar o percurso inverso, que vai da linguagem de programação em direção à linguagem de máquina, passando pela linguagem de montagem. Sob essa perspectiva, a linguagem natural traduz o sistema de operações *previsto* na memória, conforme estabelecido pela máquina, por meio de um léxico cujo espectro semântico engloba as operações realizadas na máquina abstrata. Essa tradução ocorre ao longo de um processo que converte a metalinguagem descritiva, do sistema simbólico e natural, em uma linguagem formal. Nesse percurso, no nível de montagem, os dados armazenados em uma célula de memória da máquina são convertidos por meio do empréstimo de termos oriundos da linguagem natural, a exemplo dos comandos *store*, *add* e *load*, que indicam, respectivamente, armazenar, adicionar e carregar. No nível da montagem, também é possível usar nomes simbólicos para inserir a linguagem natural no código, por exemplo, durante a rotulagem, quando significados arbitrários são usados com o objetivo de introduzir um significado linguístico que torne o código legível por humanos.

Como já pudemos mencionar, há, no nível da linguagem de programação, uma proliferação de estratégias discursivas. Elas podem ser de ordem sintática, mas também de ordem semântica, já que possibilitam atender a necessidades expressivas específicas, abrindo, desse modo, alternativas estilísticas reais. Não à toa, nesse nível, Valle e Mazzei (2017) abordam, justamente, a questão dos estilos enunciativos em diferentes linguagens de programação, o que equivale a explorar a maneira como as categorias de enunciação (pessoa, espaço, tempo) são definidas e exploradas por cada linguagem.

16 Um conjunto interessante de trabalhos pode ser encontrado nas discussões que ocorrem em torno dessa questão em comunidades de programadores em *blogs* como o Stackoverflow. Nele, os programadores distinguem entre estilos de codificação bons e ruins. A declaração a seguir aborda claramente uma questão de retórica enunciativa: "Você também deve saber que, em Python, iterar sobre índices inteiros é um e*stilo ruim* e mais lento [...]. Se você quiser apenas examinar cada um dos itens de uma lista ou ditado, faça um *loop* diretamente na lista ou no ditado". Nesse caso, vê-se que o estilo correto pode ser identificado com uma ação de processamento muito rápida. Disponível em: https://stackoverflow.com/questions/4170656/for-loop-in-python. Acesso em: abr. 2024.

17 A linguagem de programação é um actante coletivo híbrido que faz a mediação entre a linguagem natural e o referente da máquina. Em Valle (2019b), o autor estuda um trecho do Basic em que a numeração das linhas de código vai de 10 [11, 12...] a 20 e de 20 [21, 22...] a 30, deixando o espaço, entre os intervalos das linhas, reservado para adições e transformações do código por outros usuários/codificadores. Isso demonstra que escrever o código o prepara para futuras intervenções, modificações, especificações. Assim, cada código é sempre um empreendimento *in fieri* e coletivo.

As linguagens imperativas estudadas por Valle e Mazzei (2017)[18] nos servem como um bom exemplo. Do ponto de vista da pessoa, a dimensão enunciativa equivale ao comando, ao imperativo estabelecido por uma espécie de debreagem enunciativa que produz um *eu-tu*, nesse caso, dotado de um duplo modo de existência: atualizado, pois consiste em um pacote de instruções, e realizado, pois o texto com instruções é acionado e executado ao mesmo tempo.

No que diz respeito ao *espaço de enunciação*, é preciso pensar na localização da memória que, na linguagem artificial da programação, assume o lugar do espaço enunciativo de referência, isto é, do nosso próprio espaço (corporal). Nesse paradigma, tudo ocorre a partir da oposição entre uma dimensão ativa, equivalente ao ato de codificação, e uma dimensão passiva, correspondente ao próprio espaço. Quanto à *temporalidade da enunciação*, a lógica do comando não se refere a um passado ou a um futuro de ação, mas sim a uma ação que se valida em função do tempo de sua própria enunciação/execução. Dessa maneira, o tempo representado sob a forma da ordem dos enunciados se torna o tempo da enunciação em ato.

Resumidamente, podemos dizer que, no nível da linguagem de máquina, não há possibilidade de variação. Já no nível da linguagem de montagem, a variação é bastante limitada, tendo em vista que esse tipo de linguagem consiste em um *envelope linguístico* revestindo a semiótica não linguística da linguagem de máquina. Por sua vez, nas linguagens de programação, a variação é fundamental para a construção de estratégias de escrita (as chamadas "metodologias de desenvolvimento"), para a introdução de elementos normativos (guias de estilo de uma comunidade), para a articulação de sistemas de valores específicos (por exemplo, concisão *versus* clareza, clareza *versus* elegância etc.) e para a dimensão semântica – já que, especialmente no nível da montagem, recorre-se exclusivamente à sintaxe.

Banco de dados, algoritmos e criatividade: entre sedimentação e inovação

Esta segunda parte tem como objetivo entender a relação entre bancos de dados arquivados, a operação de algoritmos e a criatividade humana/máquina. Ela consiste em duas seções.

A primeira parte examinará a relação entre bancos de dados e trabalho algorítmico sob a perspectiva da análise de bancos de dados, uma questão que está no centro de nosso projeto de pesquisa atual em semiótica, visão computacional e história da arte digital, intitulado "Em direção a uma genealogia das formas visuais. Semiótica e abordagens computacionais para grandes coleções de imagens". O objetivo é analisar como a visão computacional, em conjunto com outras disciplinas, como a história da arte, pode ser usada para analisar grandes quantidades de dados usando algoritmos apropriados.

A segunda seção desta segunda parte do artigo trata dos modelos utilizados pelo Midjourney (ou mesmo pelo DALL•E, para mencionar outro exemplo) que traduzem enunciados verbais (*prompts*) em enunciados visuais, ou vice-versa: produzem enunciados verbais com o objetivo de, por exemplo, *descrever uma*

18 Os autores também levam em conta dois outros paradigmas: o *funcional* e o *orientado a objetos*.

DOI 10.5935/1980-6914/eLETTR17164

imagem que o usuário propõe ao Midjourney. Ousaríamos dizer que a geração de textos visuais por esse modelo nos interessa mais do que os experimentos com o ChatGPT, pois, especialmente no caso do Midjourney, a tradução não ocorre apenas entre a linguagem de máquina e a linguagem humana. Ela ocorre, principalmente, entre a linguagem verbal do *prompt* (o comando dado) e a linguagem visual (o produto gerado) – daí a centralidade da ferramenta de IA escolhida –, ainda que, em ambos os casos (Midjourney e ChatGPT), o princípio seja sempre o mesmo: as instruções são aplicadas a bancos de dados verbais e visuais, que desempenham um papel fundamental nessas operações de análise, tradução e produção de enunciado.

Trata-se, como veremos a seguir, de um papel que pode, em termos semióticos, ser concebido segundo a noção de enciclopédia proposta por Umberto Eco (1984) ou, ainda, em termos greimasianos e pós-greimasianos, como o *local da sedimentação de formas discursivas verbais e visuais*, pensando agora no mecanismo de renovação da cultura humana formalizado por Jacques Fontanille (1999) no esquema da práxis enunciativa. Nesse esquema, o banco de dados ocuparia o lugar da virtualização, ou seja, dos objetos culturais e dos discursos sedimentados em uma memória coletiva, e em arquivos, a partir dos quais *novas criações/performances podem ser produzidas*, nesse caso, "automaticamente", pois estamos lidando com linguagens artificiais. A teoria da práxis enunciativa nos servirá, portanto, para estudar a dinâmica *entre inovação e sedimentação* no contexto de bancos de dados entendidos como arquivos e como locais onde o novo é gerado.

Bancos de dados como sistema de cotextos: análise de semelhanças/dissimilaridades

Vamos começar com a primeira seção da segunda parte de nosso artigo.

De modo bastante sumário, poderíamos definir a IA como uma ferramenta dedicada a realizar tarefas no lugar de um humano que a treinou previamente. Ensinar uma máquina consiste, essencialmente, em capacitá-la a aprender a executar uma tarefa a partir de um banco de dados apropriado. Para isso, de início, o programador deve escolher o tipo de algoritmo de aprendizado (*random forest*, *svm* etc.), o que equivale a escolher uma estratégia segundo a tarefa a ser executada e a natureza dos dados fornecidos (imagens, planilhas etc.).

No caso da análise de grandes coleções de imagens, iremos considerar duas estratégias. A primeira, a *extração de recursos*, é a estratégia usada por Lev Manovich (2020). Ela consiste em um método que extrai recursos do conteúdo dos bancos de dados com base em regras definidas prévia e "manualmente" pelo pesquisador, que dita as instruções computacionais a serem seguidas para a execução da tarefa. É o caso, por exemplo, da escolha dos recursos a serem extraídos, como os gradientes de luminosidade em pinturas de artistas abstracionistas do início do século XX. A segunda estratégia é a *aprendizagem profunda* (*deep learning*)[19], que consiste em um algoritmo responsável por fornecer à máquina um conjunto de dados por meio dos quais e nos quais ela deve detectar semelhanças/dessemelhanças.

19 A aprendizagem profunda (ou *deep learning*) consiste em um algoritmo que define um modelo, frequentemente uma rede neural, a partir de um conjunto inicial de parâmetros, otimizando gradualmente (aprofundando) as variáveis para realizar a tarefa desejada (N. T.).

Quando usamos um algoritmo de aprendizagem profunda, não estamos mais na *extensão do olho do pesquisador* que decide o que a máquina deve encontrar na coleção de imagens (como foi o caso da extração de recursos). Trata-se de uma outra situação, que nos coloca na extensão do banco de dados usado para treinar o algoritmo. Ora, ao adotarmos a aprendizagem profunda como estratégia, estamos, na realidade, deixando o próprio algoritmo decidir sobre o que ele deve calcular para realizar satisfatoriamente a sua tarefa. Nesse caso, resta ao pesquisador apenas fornecer ao modelo um parecer sobre os resultados apresentados, permitindo que ele se corrija sem, no entanto, dizer-lhe exatamente quais cálculos deveria ter feito. De fato, *é a qualidade do banco de dados que determinará a capacidade do modelo de aprender a executar sua tarefa de forma mais ou menos correta*. Evidentemente, se o algoritmo tiver sido treinado em um conjunto de dados composto por imagens comuns, que representam objetos do cotidiano, por exemplo, será muito difícil obter bons resultados no contexto da pesquisa de imagens artísticas[20]. Dito de outro modo, *o conjunto de dados no qual o algoritmo é treinado deve ter afinidades suficientes com o banco de dados que será apresentado posteriormente*: só assim ele pode analisá-lo de modo relativamente satisfatório em termos de semelhanças e/ou diferenças.

As tarefas executadas pela máquina "em nosso lugar" – e que vimos estudando há alguns anos (Dondero, 2020) – estão relacionadas, principalmente, à análise de imagens. Obviamente, sobretudo quando o que está em jogo é a classificação de grandes coleções de dados visuais (milhares de imagens), a máquina está sendo solicitada a realizar uma tarefa que ultrapassa a capacidade puramente humana de análise. Nesse caso, é importante destacar que estamos falando de dados *digitalizados*, e não de dados *digitais*: pretendemos nos concentrar na análise de grandes coleções de imagens, pertencentes ao patrimônio artístico ocidental, e não nos dados produzidos pela própria tecnologia digital, a exemplo daqueles que geramos cotidianamente nas redes sociais, *e-mails* etc.

Logo, podemos notar que a produção desses dados massivos possibilitou a análise de grandes coleções de imagens, reabrindo, inclusive, o terreno para projetos de pesquisa que antes não eram sequer conjecturados. Referimo-nos, aqui, em particular, aos projetos de dois historiadores da arte: Henri Focillon e, sobretudo, Aby Warburg. Em seu livro seminal *Vie de formes*, Focillon (1934) teorizou uma genealogia das formas como um processo de bifurcações/ramificações/estratificações/desaparecimentos de formas, sem, no entanto, poder seguir uma genealogia em sua totalidade. Já Warburg (2012) estudou a imagem por meio da imagem, usando como método de investigação a visualização, de modo que imagens com características comuns fossem dispostas próximas umas das outras em grandes painéis pretos, em função de semelhanças de composição.

Diante das múltiplas questões que cada imagem artística coloca para o observador e para o historiador, Warburg (2012) escolheu como resposta e como explicação geral uma fórmula, por assim dizer: aproximar uma imagem de "sua vizinha mais próxima", em termos plásticos, e distanciá-la daquelas às quais ela

20 Em geral, começa-se compilando o conjunto completo de dados que desejamos analisar. Em seguida, extraímos parte dele, fazemos anotações e o usamos para treinamento. O restante será utilizado para verificar os resultados, que são, de fato, o que queremos estudar. Se essa distribuição do conjunto de dados inicial for bem-sucedida, teremos mais chances de ter um modelo que generalize mais precisamente os dados de interesse.

DOI 10.5935/1980-6914/eLETTR17164

se opõe ou com as quais entra em conflito, gradualmente. Foi exatamente isso que fizeram os pesquisadores que, a exemplo de Lev Manovich, seguiram essa proposta, garantindo que os bancos de dados e os seus algoritmos pudessem agrupar dados de acordo com suas semelhanças e diferenças, segundo a regra do "vizinho mais próximo", de Warburg.

Figura 2 – Aby Warburg, Atlas Mnemosyne, 1924-1929. Bilderatlas Panel 45
Fonte: https://warburg.sas.ac.uk/archive/bilderatlas-mnemosyne

Há alguns anos, estudamos dois modelos de visualização (Dondero, 2017, 2019b, 2020) que exemplificam bem as possibilidades dessa proposta. O primeiro é uma montagem clássica, com cerca de 4.500 imagens (Figura 3). Já o segundo, o mais interessante, são as visualizações que chamamos de diagrama de imagens (Figura 4). A montagem nos parece ser menos interessante por um motivo talvez evidente: ela segue uma organização determinada por um metadado, nesse caso, a data de produção[21]. Já no caso do diagrama, a disposição das imagens depende unicamente das instruções que o pesquisador dá à máquina – e não dos metadados, como ocorre com a montagem. Essas instruções têm como objetivo medir a semelhança visual entre as características plásticas das imagens contidas no banco de dados[22].

21 Trata-se de uma estratégia que já criticamos em várias publicações (Dondero, 2017, 2019b, 2020), nas quais explicamos que a organização de coleções por meio de metadados recai no mesmo erro pelo qual se critica Roland Barthes em relação à translinguística, ou seja, a tentativa de reduzir a imagem ao que pode ser lexicalizado. Disponível em: https://www.academia.edu/34822560/Barthes_entre_sémiologie_et_sémiotique_le_cas_de_la_photographie_2017_Roland_Barthes_Continuités_Cerisy_2016_J_P_Bertrand_dir_. Acesso em: abr. 2024.

22 Para uma análise enunciativa desses dois tipos de visualização de imagem, ver o terceiro capítulo de *The language of images: The Forms and the Forces* (Dondero, 2020), em que fazemos a distinção entre as focalizações relevantes para a montagem e aquelas relevantes para os diagramas, adotando, para isso, a classificação de Fontanille (1998) proposta em *Sémiotique et littérature*.

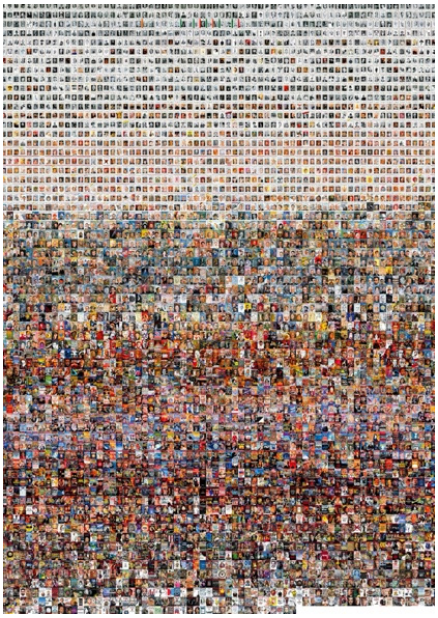

Figura 3 – Montagem de imagens. 4.535 Capas da *Time* Magazine (1923-2009). Manovich & Douglass (2009)

Fonte: https://manovich.net/

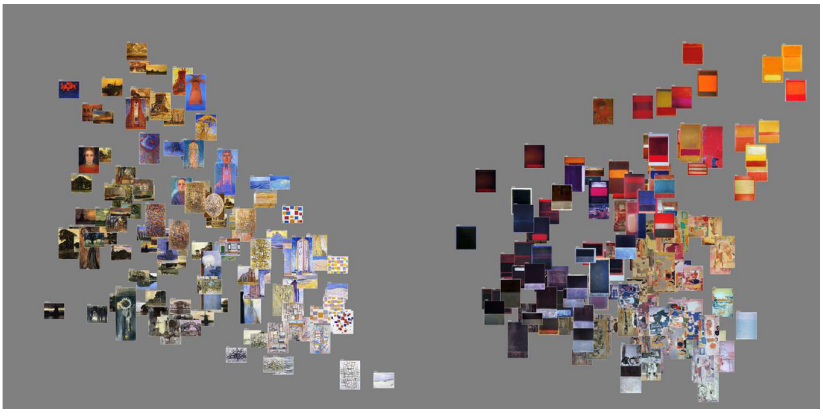

Figura 4 – Comparação de 128 pinturas de Piet Mondrian (1905-1917) e 151 pinturas de Mark Rothko (1944-1957). As duas visualizações de imagem são colocadas lado a lado, de modo que compartilham o mesmo eixo X. Eixo X: média de brilho. Eixo Y: média de saturação. Manovich, Douglass, Zepel (2011)

Fonte: https://manovich.net/

Dentre as propriedades plásticas, a categoria cromática é a mais fácil de ser trabalhada pela máquina porque é uma categoria comensurável, assim como as intensidades de luz. De fato, no caso do método de "extração de recursos" de que falamos aqui, o objetivo é extrair das imagens as características plásticas que,

DOI 10.5935/1980-6914/eLETTR17164

na visão computacional, são chamadas de "recursos de baixo nível"[23]. Trata-se de propriedades que não estão diretamente ligadas à figuração. Entretanto, essa tarefa diz respeito à aferição das médias de cada característica distribuída na superfície de cada imagem, e não se confunde, portanto, com a identificação de *formantes plásticos* ou *figurativos* (Greimas, 1984).

Em alguns trabalhos que publicamos anteriormente (Dondero, 2017, 2019b, 2020), fizemos, justamente, uma crítica a essa metodologia de análise de grandes coleções de imagens: o procedimento de extração trabalha com características plásticas *médias*[24], deixando de se concentrar na *distribuição* dessas características dentro da imagem artística, entendida como uma totalidade[25]. Mas, a despeito das críticas que podem ser feitas a este ou àquele método estatístico, há inúmeros motivos que justificam o interesse semiótico nessas análises, dentre os quais listamos dois:

- Além de desenvolverem uma das questões de Warburg (a das imagens e de seus "vizinhos mais próximos"), essas visualizações também revelam um trabalho que pode ser entendido como estruturalista porque contrasta grupos de imagens com características plásticas gradualmente semelhantes ou opostas e organizam as características de cada imagem *gradualmente*, dentro de um espaço de controle (uma perspectiva tensiva da estrutura).
- Essas visualizações de imagens apresentam a análise realizada (no sentido de divisão, agrupamento, reconstrução da relação) e permitem efetuar um raciocínio diagramático, colocando em jogo aspectos estatísticos e perceptuais via *semissimbolismo*, de acordo com o parâmetro de similaridade/dissimilaridade. Podemos usar essas visualizações para realizar experimentos estatístico-perceptuais em uma coleção com base em vários parâmetros relevantes para cada banco de dados (que não se limitam a características cromáticas ou luminosas, pois incluem, também, aspectos da geometria das formas, do comprimento e da tipologia das linhas desenhadas). Isso não é possível com as visualizações de tendências[26], já que elas se constroem por símbolos ou números.

Como já pudemos indicar mais de uma vez, o sistema de coleção funciona como uma enciclopédia, em outras palavras, como um sistema de cotextos, para usar um termo de Umberto Eco (1984). Ou, ainda, como o lugar onde as estratégias discursivas de formas artísticas (por exemplo, formas pictóricas) de todos os tempos foram sedimentadas. Portanto, podemos situá-las no diagrama da práxis enunciativa *no espaço da virtualização*.

23 A expressão se refere a características básicas e primitivas das imagens, extraídas sem necessidade de uma modulação semântica, que podem incluir cores, texturas, bordas, formas, histogramas, gradientes, entre outros (N. T.).

24 Há, no entanto, uma observação: quando usamos uma rede como a ResNet para extrair/calcular uma incorporação de uma imagem, que chamamos de extração de recursos (pelo menos, no caso da visão computacional), essa incorporação ainda contém informações relacionadas à distribuição de recursos e não apenas à média.

25 Sobre a imagem como um todo, ver Goodman (1976), Thom (1983) e Dondero (2020).

26 Visualizações de tendência são um tipo de visualização de imagem no qual elas se organizam segundo o desenvolvimento de padrões ou direções em conjunto de dados visuais, geralmente ao longo do tempo (N. T.).

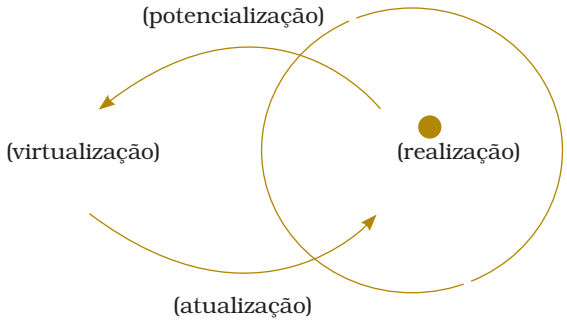

Figura 5 – Diagrama da práxis enunciativa
Fonte: Fontanille (2015, p. 275).

Muitas das perguntas feitas na semiologia visual e na semiótica, desde a década de 1960, ainda não foram resolvidas – por exemplo: "Existe uma linguagem visual?". No entanto, ao menos essas questões foram postas concretamente e, em parte, respondidas graças às operações de algoritmos realizadas em bancos de dados de imagens. De fato, um banco de dados não é propriamente equiparável à *linguagem* saussuriana, que é composta de virtualidades, mas ele tem uma espessura conceitual muito semelhante à do *locus* da virtualização: as imagens que este contém foram produzidas historicamente e, de certa forma, estão copresentes no banco de dados. Não se trata, portanto, de meras virtualidades pictóricas, mas de textos atestados, depositados no banco de dados e que compartilham, após a digitalização, uma substância comum e responsável por torná-los *comensuráveis* e disponíveis para análise algorítmica. Essa comensurabilidade significa que, em um banco de dados, as imagens podem ser manipuladas e medidas até que sua especificidade/diferença se destaque das demais.

Pensando, ainda, na proposta de Warburg (2012), o nosso projeto de pesquisa *Towards a genealogy of visual forms: semiotics and computational approaches to large collections of images* (ver nota 8) é outra maneira – mais complexa, esperamos – de dar continuidade à genealogia das formas, em particular às formas de *páthos* (o *pathosformeln*[27] de Warburg) que podemos ver nas poses e nos gestos das figuras retratadas em pinturas ao longo da história. Mas não se trata apenas disso. Trata-se também de fazer avançar o próprio projeto da semiótica visual, em particular, uma questão específica que já abordamos (cf. Dondero, 2017, 2019b), sobretudo em nosso livro *The language of images* (Dondero, 2020): o estudo do movimento, da temporalidade e da narratividade na imagem estática, a partir do modo como a enunciação temporal e aspectual é significada em uma substância fixa como a pintura.

Ora, a enunciação da categoria de pessoa foi amplamente desenvolvida e estudada em trabalhos sobre o rosto e o perfil (Beyaert-Geslin, 2017; Dondero, 2023, 2024). Isso também ocorreu com a enunciação espacial – nesse último caso, em especial, graças a vários trabalhos sobre perspectiva, tais como os de Marin (1993) e Fontanille (1989). Todavia, a enunciação temporal (o antes e o

27 Conceito cuja tradução literal seria "formas de *páthos*". Diz respeito ao sentimento veiculado (culturalmente) por gestos, poses e expressões faciais. No contexto da visão computacional, em que se insere a autora do artigo, o tratamento da pose é crucial para o aprimoramento da interação ser humano-máquina, por exemplo (N. T.).

DOI 10.5935/1980-6914/eLETTR17164

depois dentro de uma ação representada em imagem), a enunciação aspectual (o momento da ação, focalizado pelo produtor) e o ritmo do desenrolar da ação não foram suficientemente investigados[28].

Podemos tomar alguns exemplos de um *corpus* produzido a partir da coleção completa de pinturas disponíveis no WikiArt — triadas, evidentemente, de acordo com nossas necessidades. Chegamos a um grupo de cinco mil imagens religiosas, a partir do qual buscamos formalizar gestos/poses: primeiro individuais e depois coletivos. Usamos o MMPose[29] para mapeamento e o PixPlot[30] para visualização. Por que poses? Porque são o local de uma dinâmica figurativa, ou seja, de um *continuum*, e porque nosso desafio é estudar, justamente, o *movimento em imagens estáticas de maneira automática*. No MMPose, as poses são mapeadas a partir de 17 pontos-chave, que tentam abranger todo o corpo. Quando os 17 pontos não foram contemplados, excluímos a imagem do *corpus*, pois o algoritmo não identifica um corpo inteiro. Na Figura 6, é possível acompanhar essa organização.

Figura 6 – Visualização das poses no *corpus* de trabalho extraído do WikiArt

Fonte: Adrien Deliège e Maria Giulia Dondero. Visualização gerada via MMPose e o PixPlot.

O círculo vermelho, no diagrama geral da coleção, à esquerda, relaciona-se a um grupo de pinturas em que a máquina reconhece uma pose específica. Podemos percorrer toda a coleção – que chamamos de *corpus de referência*, por ser abrangente – e selecionar o *corpus de trabalho* – também segundo nossa terminologia –, composto por vários grupos de poses em comparação (circuladas em vermelho). Elas podem, por exemplo, ser usadas em pesquisas sobre a relação entre a expressão dessas imagens e seu conteúdo, entre outras. Ainda há várias questões a serem consideradas: precisamos decidir se criaremos uma genealogia

28 Há pouquíssimos estudos sobre essa questão: poderíamos mencionar, por exemplo, o trabalho de Petitot (2004) sobre o *Laocoon*, um artigo do Groupe μ (1998) e outro de Colas-Blaise (2019).

29 O MMPose é um *framework* (conjunto de ferramentas, bibliotecas e convenções que proporcionam uma base para o desenvolvimento de um *software*) de código aberto, destinado a mapear poses humanas em imagens e vídeos, por meio do rastreamento da angulação e direção articulares, como se vê na Figura 6 (N. T.).

30 O PixPlot é uma ferramenta de visualização interativa desenvolvida para explorar grandes coleções de imagens de modo bidimensional. Ele permite identificar padrões visuais, agrupamentos e relações entre imagens de maneira intuitiva ou automática (N. T.).

das poses coletivas mais semelhantes ou das *formas de poses* mais semelhantes (algumas poses coletivas podem formar triângulos, quadrados etc.) ou ainda se levaremos em conta o local das poses e sua escala na superfície da imagem[31].

Geração de novas imagens

Até aqui, fizemos um percurso em que saímos das teorias da enunciação, passamos pela análise computacional de *big data* e chegamos, finalmente, à *geração automática e algorítmica de novas imagens*. Neste último caso, mais uma vez, é a máquina que enuncia por meio de nossas instruções, traduzindo-as da linguagem verbal para a linguagem visual. Mas ela também pode fazer o oposto, como já dissemos: descrever uma imagem em linguagem verbal. Nesse aspecto, as possibilidades enunciativas de geradores como o Midjourney são bastante amplas para cada direção de tradução (verbal ↔ visual): é possível mudar o estilo de uma foto, misturar vários deles ou fundir imagens de artistas diferentes. No entanto, cada nova imagem começa com outro tipo de tradução, mais fundamental, como assinalamos anteriormente: a tradução da imagem e do texto verbal em números, que gera listas numéricas conhecidas como *embeddings*. A capacidade de manipulação da imagem, adquirida dessa forma, permite que a IA generativa produza novas imagens automaticamente, usando bancos de dados e métodos de aprendizado de máquina. Essas novas imagens são geradas por operações realizadas em todas as imagens já produzidas, que são armazenadas e anotadas de acordo com o estilo, o autor e o gênero nos bancos de dados disponíveis (WikiArt, Artsy, Google Art and Culture etc.).

Os modelos de geração de imagens usam um componente *Large Language Models*[32], ou pelo menos um modelo que entenda a linguagem natural (por exemplo, *Contrastive Language-Image Pre-training* – Clip[33]), e que transforma *prompts* (comandos) em *embeddings* (listas de números) que podem ser usados pela máquina[34]. As listas de números que descrevem imagens são vinculadas a listas de números que identificam textos em linguagem natural. Esses modelos de aprendizado, responsáveis pela tradução entre linguagens verbais e visuais, são determinados pela organização do conteúdo do banco de dados. Entretanto, atualmente pode-se produzir imagens a partir de dados digitalizados que remetem a todos os estilos e aos artistas mais famosos da contemporaneidade[35]. As imagens geradas por computador nos permitem entender como grandes obras de artistas do passado podem ser misturadas e, em alguns casos, apontar os

31 Esse projeto está em andamento e o restante dessa pesquisa pode ser encontrado em Dondero (2025).

32 Sistemas de IA projetados para entender e gerar linguagem natural, treinados em grandes quantidades de texto para aprender padrões linguísticos, permitindo que se execute uma variedade de tarefas relacionadas ao processamento de linguagem, como tradução automática, geração de texto, resumo de documentos, resposta a perguntas etc. Os mais conhecidos são o ChatGPT da OpenAI e o BERT da Google (N. T.).

33 Modelo de IA desenvolvido pela OpenAI projetado para entender tanto texto quanto imagens *de maneira integrada*. Isso permite que o modelo associe palavras e frases com conteúdo visual, sem a necessidade de treinamento supervisionado (N. T.).

34 Isso no caso de modelos para gerar textos verbais, como GPT-3.5, GPT-4, Llama, Claude, PaLM.

35 Algumas imagens produzidas por máquinas chegaram até mesmo a ganhar prêmios em competições de imagens produzidas por humanos. Esse é o caso da "Feira Estadual do Colorado", em que Jason Allen venceu o concurso de arte graças ao seu trabalho intitulado *Théâtre d'Opéra Spatial*, produzido com o Midjourney. Disponível em: https://www.tomsguide.fr/polemique-une-oeuvre-dart-generee-par-lia-remporte-un-concours-les-artistes-sindignent/. Outro exemplo de uma obra de arte produzida por IA é *Unsupervised*, 2022, de Refik Anadol Studio. Esse projeto usa redes neurais treinadas em um banco de dados de dez mil obras de arte da coleção do Museum of Modern Art (MoMA). Essa coleção inclui arte de 1870 a 1970, bem como obras de décadas posteriores. Sobre esse tema, indicamos o trabalho de Manovich e Arielli (2021-2024) e o paradoxo que ele destaca haver entre o movimento do modernismo – ao qual as obras do banco de dados pertencem, que visa ao novo e à destruição do antigo – e os algoritmos que as retrabalham.

DOI 10.5935/1980-6914/eLETTR17164

estereótipos mais comuns de cada artista ou movimento artístico. Vejamos, por exemplo, os estereótipos que a máquina nos apresentou a partir dos extensos bancos de dados sobre os estilos do Renascimento, Barroco, Maneirismo e Rococó (Figura 7).

Figura 7 – *Prompt: /Ascensão de Maria Madalena* nos estilos da Renascença, do Barroco, do Maneirismo e do Rococó/

Fonte: Experimento realizado por M. G. Dondero, Enzo D'Armenio e Adrien Deliège via Midjourney (2023).

Também pedimos ao Midjourney que gerasse imagens estereotipadas de Van Gogh, por meio do *prompt*: /uma paisagem no estilo de Van Gogh/ (ver Figura 8). Rapidamente percebemos que seria difícil livrar-nos de determinados objetos, sobretudo o sol.

Figura 8 – *Prompt*: /uma paisagem no estilo de Van Gogh/

Fonte: Experimento realizado por M. G. Dondero, Enzo D'Armenio e Adrien Deliège via Midjourney (2023).

Isso ocorreu porque, provavelmente, essa figura é considerada um recurso predominante na obra de Van Gogh (a depender das correspondências entre as incorporações das imagens e das descrições das imagens que foram codificadas). Uma primeira tentativa, talvez "ingênua", de fazer o sol desaparecer foi adicionar /*without sun without moon/* (sem sol) ao *prompt* (ver Figura 9).

Figura 9 – *Prompt*: /uma paisagem no estilo de Van Gogh without sun without moon/

Fonte: Experimento realizado por M. G. Dondero, Enzo D'Armenio e Adrien Deliège via Midjourney (2023).

Podemos ver que as imagens mantêm o sol (ou uma lua?, é difícil dizer), pois o Midjourney não foi projetado para "pensar" realmente sobre o significado do *prompt*, nem para distinguir entre os significados positivo e negativo das nossas solicitações. Conforme indicado na documentação do Midjourney, de fato, uma palavra que aparece no *prompt* tem mais probabilidade de ser representada na imagem. Descobrimos que para se livrar de um elemento, o usuário precisa usar o comando especial "- -" (*minus minus*) /*sun, moon*/ (Figura 10).

Figura 10 – M. G. Dondero, E. D'Armenio, A. Deliège, Midjourney. *Prompt*: "Landscape in Van Gogh Style" - - sun, moon, 2023.

Fonte: Experimento realizado por M. G. Dondero, Enzo D'Armenio e Adrien Deliège via Midjourney (2023).

DOI 10.5935/1980-6914/eLETTR17164

Consideramos esses exemplos muito significativos porque entendemos que a negação em imagens é produzida exclusivamente por ir além do *prompt* e do nível de tradução que a máquina pode fornecer atualmente entre o *prompt* e a forma visual. Portanto, precisamos usar comandos que nos permitam agir diretamente na imagem sem passar pelo processo de tradução do *prompt*.

Ao trabalhar com essa ferramenta de IA, percebemos que a produção de uma série de imagens exige que o utilizador execute várias operações[36]. Quando uma instrução é dada ao Midjourney, são obtidas, por padrão, quatro versões dessa instrução verbal, que diferem entre si em termos de intensidade da luz, posicionamento dos objetos etc. O experimentador deve escolher a melhor e decidir – ou não – continuar buscando a imagem ideal, dando instruções adicionais para modificar a versão escolhida. É possível transformar as quatro versões produzidas (que podem ser entendidas como diferentes *otimizações da instrução* dada) escolhendo uma em cada série de quatro, até que o resultado corresponda à imagem almejada pelo experimentador[37].

O experimentador, se for um programador, pode decidir refinar (ajustar) uma rede neural por meio de anotações, construindo correspondências mais precisas entre as listas de números que identificam as descrições em linguagem natural e as listas de números que identificam as imagens. De nossa parte, para tornar a produção de imagens mais próxima de nossos objetivos e, assim, minimizar o viés ou o ruído gerado por bancos de dados excessivamente genéricos, podemos, no máximo, refinar nosso *prompt* fornecendo-lhe mais indicações[38].

Outro modo de criar restrições que limitam a generalidade dos resultados é indicar explicitamente a técnica a ser usada, como /desenho a giz/, /pintura a óleo/, /afresco/ etc., além, é claro, de precisar um ou mais estilos pictóricos[39]. No nosso caso, o que é particularmente importante é testar a mistura de diferentes estilos de pintores de acordo com suas características e refletir sobre várias situações interessantes que surgem com relação à composição. Os experimentos de Lev Manovich (publicados no Facebook em 2023) e em Manovich e Emanuele Arielli (2021-2024) são bastante convenientes nesse sentido: neles, as figuras de Bosch mudam de acordo com as posições ocupadas na paisagem, cujas coordenadas são dadas por padrões geométricos inspirados em Malevich (Figura 11).

Se observarmos outra produção de Manovich e Arielli (2021-2024), que mistura Brueghel e Kandinsky, parece-nos ser possível argumentar que artistas abstracionistas como Malevich e Kandinsky são usados pela máquina como paisagistas. Como se vê a seguir, eles acabam determinando a topologia geral da imagem, que acomoda as figuras de pintores como Brueghel e Kandinsky (Figura 12), tradicionalmente considerados paisagistas. Há, portanto, uma inversão de papéis.

36 Enzo D'Armênio e Adrien Deliège foram atores importantes no desenvolvimento dessas reflexões, tendo participado ativamente dos experimentos.

37 Aluminé Rosso sugeriu a George Legrady, no simpósio *Face it*!, realizado na Université de Liège de 25 a 27 de janeiro de 2023, sob organização de M. G. Dondero, M. Leone e C. Paolucci, considerar essas operações como comparáveis às de um curador de exposições, que aconselha o artista (a máquina), testa suas propostas e o ajuda a definir a versão final de seu trabalho de acordo com o ambiente de implementação.

38 O Midjourney também introduziu recentemente uma ferramenta para modificar apenas uma parte ou região da imagem produzida, previamente selecionada pelo experimentador (*vary region*): basta circular a parte a ser modificada e inserir um *prompt* que atenda às necessidades do experimentador. Trata-se de um avanço, pois as modificações por esse comando são muito mais eficientes do que modificar diretamente um *prompt*, é claro, se o que se busca é realizar modificações localizadas. Por exemplo, se já tivermos gerado um homem segurando uma raquete de pingue-pongue na mão esquerda e quisermos que ele segure uma raquete de tênis, será (na nossa opinião, mas é passível de teste) mais eficiente usar a nova funcionalidade selecionando a raquete e inserindo o *prompt* /raquete de tênis/, do que refazer um *prompt* inteiro que especifique tudo isso. Além disso, refazer um *prompt* completo poderia modificar a imagem mais do que o desejado.

39 Essa dimensão é aparentemente muito importante para a máquina, mas foi relativamente ignorada na semiótica até os desenvolvimentos teórico-metodológicos relativo aos suportes.

Figura 11 – Experimento com o *prompt*: /pintado por Malevich e Bosch/

Fonte: Manovich e Arielli (2021-2024) via Midjourney.

Figura 12 – Experimento com o *prompt*: /pintado por Brueghel e Kandinsky/

Fonte: Manovich e Arielli (2021-2024) via Midjourney.

Também buscamos misturar alguns estilos pictóricos. Os resultados são frustrantes e, em alguns casos, divertidos. Um exemplo é a mistura entre Da Vinci e Rothko (Figura 13). Esses dois pintores, separados por alguns séculos, foram reconhecidos como especialistas em perspectiva atmosférica e em contornos imprecisos e camadas de cor, respectivamente. Alguns dos resultados foram decepcionantes, por exemplo: a máquina nos forneceu uma *Mona Lisa* sobreposta banalmente por um triângulo vermelho-Rothko. Todavia, obtemos resultados mais interessantes quando os estratos de cor de Rothko, por vezes beirando a transparência, apareceram sobrepostos à perspectiva atmosférica de Da Vinci.

DOI 10.5935/1980-6914/eLETTR17164

Figura 13 – *Upload* da *Mona Lisa* de Leonardo da Vinci no Midjourney + *prompt* /estilo Rothko/

Fonte: Experimento por M. G. Dondero, Enzo D'Armenio e Adrien Deliège via Midjourney (2024).

Nessas quatro imagens, podemos notar que a adição de desfoque e de transparência transforma a paisagem de Da Vinci: de desfocada (em função da distância imposta pela perspectiva) em nítida, lembrando, neste último caso, as pinturas americanas hiper-realistas da década de 1970. Considerando todos esses experimentos, resta-nos o seguinte questionamento: "O Midjourney é programado para atingir, sempre, um equilíbrio entre o desfocado e o nítido, o impreciso e o detalhado?". Em última instância, vemos que é somente por meio da produção de uma infinidade de imagens, da mistura de estilos e de técnicas de produção – ou seja, somente reiterando nossas solicitações – que seremos capazes de compreender o espaço de linguagem/virtualização que está por trás dessas produções. É a partir de uma infinidade de imagens geradas automaticamente que estaremos aptos a construir hipóteses sobre o banco de dados no qual Midjourney foi treinado e, portanto, sobre seu modelo (mantido sob segredo).

Algumas considerações parciais

Chegamos a certas conclusões que são, ainda, bastante provisórias. Do ponto de vista da enunciação enunciada, ou seja, da maneira como o ato de produção se reflete no enunciado produzido, o Midjourney é capaz de usar um estilo para cada pintor e para cada pintura que se quer produzir. No caso de Van Gogh, por exemplo, o Midjourney usa a textura típica do pintor e imita uma motricidade sensorial que é bastante semelhante ao ritmo de seu toque. No entanto, a máquina tem seu próprio estilo, ou seja, uma opacidade enunciativa da mão[40], que se aproxima, a nosso ver, do estilo do expressionismo pictórico americano dos anos 1970.

40 Ver Marin (1993).

Do ponto de vista da práxis enunciativa, do mecanismo formal de renovação e alimentação das culturas, esse processo é operacionalizado por meio de seus modos de existência. O banco de dados desempenha o processo de virtualidade/virtualização das formas, pois as imagens dos pintores que ele contém, no caso do Midjourney, podem ser vistas como formas sedimentadas de nossa cultura visual ocidental. Já os procedimentos que desencadeamos via *prompts* podem ser vistos como uma etapa de atualização realizada nas imagens geradas. No que diz respeito à potencialização, as palavras que produzimos, ou seja, os enunciados gerados por meio de nossos *prompts*, não serão – talvez nunca – imediatamente sedimentadas e aceitas no banco de dados, que é estabilizado, fixo. Afinal, teríamos de nos tornar artistas reconhecidos para podermos realizar esse feito e, assim, participar da transformação daquilo que está sedimentado nos bancos de dados e, por extensão, na própria cultura.

Artificial intelligence and enunciation: analysis of big visual data and automatic generation via Midjourney

Abstract: The AIs have been simulating, increasingly satisfactorily, the specificity of human language and actions. Therefore, it is essential for semiotics to address these technologies and practices of automation. In order to undertake this approach, starting from a diverse corpus, we will adopt the viewpoint of Émile Benveniste's theory of enunciation, focusing particularly on three developments in semiotics applied to the study of AI, which will allow us to discuss: 1. the modes of addressing the machine to the Midjourney user; 2. the translation between artificial language and human language; and 3. the relationship between databases, algorithms, and machine creativity, through the analysis of large collections of images (big visual data).

Keywords: Big visual data. Enunciation. Automatic generation of images. Artificial intelligence. Midjourney.

Referências

BENVENISTE, É. *Problems in general linguistics*. Miami: University of Miami Press, 1971.

BEYAERT-GESLIN, A. *Sémiotique du portrait*: De Dibutade au selfie. Louvain-la-Neuve: De Boeck Supérieur, 2017.

COLAS-BLAISE, M. Comment penser la narrativité dans l'image fixe? La "composition cinétique" chez Paul Klee. *Pratiques*, n. 181-182, p. 1-15, 2019. Disponível em: http://journals.openedition.org/pratiques/6097. Acesso em: jun. 2024.

DONDERO, M. G. The semiotics of design in media visualization: mereology and observation strategies. *Information Design Journal*, v. 23, n. 2, p. 208-218, 2017. DOI: https://doi.org/10.1075/idj.23.2.09don.

DONDERO, M. G. Le travail des algorithmes. Quelques réflexions sur l'actantialité et l'énonciation. *In*: CONFERÊNCIA DA AFS, 10 jun. 2019a. Lyon, Université de Lyon 2, org. Pierluigi BASSO. Disponível em: https://core.ac.uk/outputs/220155468/. Acesso em: jun. 2024.

DONDERO, M. G. Visual semiotics and automatic analysis of images from the Cultural Analytics Lab: how can quantitative and qualitative analysis be combined? *Semiotica*, v. 230, p. 121-142, 2019b. DOI: https://doi.org/10.1515/sem-2018-0104.

DONDERO, M. G. *The language of images*. The forms and the forces. Cham: Springer, 2020.

DONDERO, M. G. Emerging faces: the figure-ground relation from renaissance painting to deepfakes. *In*: LEONE, M. (org.). *The hybrid face*: paradoxes of the visage in the digital era. London: Routledge, 2023. p. 74-86.

DONDERO, M. G. The face: between background, enunciative temporality and status. *Reti, Saperi, Linguaggi: The Italian Journal of Cognitive Sciences*, 1, a. 13 (25), p. 49-70, 2024.

DONDERO, M. G. Semiótica da inteligência artificial: análise computacional de grandes bases de dados e geração automática de imagens. *Matrizes*, 2025.

ECO, U. *Semiotica e filosofia del linguaggio*. Torino: Einaudi, 1984.

FOCILLON, H. *Vie de formes*. Paris: Presses Universitaires de France, 1934.

FONTANILLE, J. *Les espaces subjectifs*. Introduction à la sémiotique de l'observateur. Paris: Hachette, 1989.

FONTANILLE, J. *Sémiotique et littérature*. Essais de méthode. Paris: Presses Universitaires de France, 1998.

FONTANILLE, J. *Sémiotique du discours*. Limoges: Presses Universitaires de Limoges, 1999.

FONTANILLE, J. *Semiótica do discurso*. Tradução Jean Cristtus Portela. São Paulo: Contexto, 2015.

GOODMAN, N. *Languages of art*: an approach to a theory of symbols. Indianapolis: Hackett Publishing Company, 1976.

GREIMAS, A.J. Sémiotique figurative et sémiotique plastique. *Actes sémiotiques*. Document VI, 1984.

GREIMAS, A. J. *Du sens II*. Paris: Seuil, 1983.

GROUPE μ. L'effet de temporalité dans les images fixes. *Texte*, n. 21-22, p. 41-69, 1998.

KEMENY, J. G.; KURTZ, T. E. *Introduction to Basic*: programming language. New York: Wiley, 1964.

KERNIGHAN, B. W.; PLAUGER, P. J. *Elements of programming style*. 2. ed. New York: McGraw-Hill, 1978.

KNUTH, D. E. *The TeXbook*. Reading: Addison-Wesley, 1984.

LASSÈGUE, J. *Turing*. Tradução Guilherme J. F. Teixeira. São Paulo: Estação Liberdade, 2017.

LIEN, J. *The Basic handbook*. New York: McGraw-Hill, 1981.

LOPES, C. V. *Exercises in programming styles*. Boca Raton: CRC Press, 2014.

MANOVICH, L. *Cultural analytics*. Cambridge: MIT Press, 2020.

MANOVICH, L.; ARIELLI, M. Artificial aesthetics: generative AI, art and visual media. 2021-2024. Disponível em: http://manovich.net/index.php/projects/artificial-aest hetics. Acesso em: jan. 2024.

MARIN, L. *De la représentation*. Paris: Seuil, 1993.

PETITOT, J. *Morphologie et esthétique*. Paris: Maisonneuve et Larose, 2004.

QUENEAU, R. *Exercises in style*. Paris: Gallimard, 1947.

TANAKA-ISHII, K. *Semiotics of programming*. Cambridge: Cambridge University Press, 2010.

THOM, R. Local et global dans l'œuvre d'art. *Le Débat*, v. 2, n. 24, p. 73-89, 1983.

TURING, A. M. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, v. 45, p. 230-265, 1936.

TURING, A. M. Computing machinery and intelligence. *Mind*, v. 59, n. 236, p. 433-460, 1950.

VALLE, A. Entrevista concedida a Pierluigi Basso e Maria Giulia Dondero. *In*: BASSO, P.; COLAS-BLAISE, M.; DONDERO, M. G. (org.). La communication à l'épreuve du geste numérique. *MEI, Médiation et Information*, n. 49, p. 45-56, 2019a.

VALLE, A. On a fragment of Basic code in Foucault's Pendulum by Umberto Eco. *Lexia*, n. 32, 2019b.

VALLE, A.; MAZZEI, A. Sapir-Whorf vs. Boas-Jakobson. Enunciation and the semiotics of programming languages. ACADEMIA, 2017. Disponível em: https://www.academia.edu/36681676/Sapir_Whorf_vs_Boas_Jakobson_Enunciation_and_the_Semiotics_of_Programming_Languages. Acesso em: mar. 2024.

WARBURG, A. *L'atlas Mnémosyne*. Paris: Éditions Atelier de l'écarquillé, 2012.