# Semiótica da inteligência artificial: análise computacional de grandes bases de dados e geração automática de imagens[a, b]

## *Semiotics of artificial intelligence: a computational analysis of big datasets and automatic image generation*

MARIA GIULIA DONDERO[c]
Fonds National de la Recherche Scientifique. Liège, Bélgica

GUSTAVO H. R. DE CASTRO[d]
Universidade Estadual Paulista. Araraquara – SP, Brasil

MATHEUS NOGUEIRA SCHWARTZMANN[e]
Universidade Estadual Paulista. Araraquara – SP, Brasil

DANIERVELIN PEREIRA[f]
Universidade Federal de Minas Gerais. Belo Horizonte – MG, Brasil

**RESUMO**

A inteligência artificial simula hoje, de maneira cada vez mais satisfatória, a complexidade da linguagem e das ações humanas. Neste artigo abordamos as inteligências artificiais com instrumentos semióticos. Aqui, assumiremos o ponto de vista da teoria da enunciação de É. Benveniste, especialmente dos desenvolvimentos em semiótica pós-greimasiana, e sobretudo de Jacques Fontanille aplicados ao estudo da inteligência artificial. Essa base teórica nos possibilitará discutir, primeiramente, a relação entre banco de dados de imagens e algoritmos na análise de grandes coleções de imagens por meio da *computer vision*, além dos modos de diálogo do usuário com o modelo de inteligência artificial generativa Midjourney, que nos permitirá tratar a criatividade da máquina.

**Palavras-chave:** Inteligência artificial, modelo generativo, práxis enunciativa, geração de imagens, Midjourney

[a] Este artigo é parcialmente baseado na pesquisa apresentada no artigo "Inteligência artificial e enunciação: análise de grandes coleções de imagens e geração automática via Midjourney" (Dondero et al., 2024). Essa pesquisa é desenvolvida, atualizada e aperfeiçoada aqui.

[b] Agradecemos a A. Deliège por suas explicações sobre métodos de análise em aprendizado de máquina. Também agradecemos a Enzo D'Armenio por seus comentários sobre a geração de imagens usando o Midjourney.

[c] Diretora de pesquisa do Fonds National de la Recherche Scientifique (F.R.S. – FNRS), na Bélgica, e professora da Universidade de Liège. Orcid: [0000-0003-2320-8130]. E-mail: mariagiulia.dondero@uliege.be

[d] Doutorando (Fapesp 2019/27000-7) do Programa de Pós-graduação em Linguística e Língua Portuguesa da Universidade Estadual Paulista (Unesp). Tradução, revisão e notas. Orcid: 0000-0003-4486-9579. E-mail: g.castro@unesp.br

[e] Livre-docente em Semiótica do Discurso pela Universidade Estadual Paulista (Unesp), atualmente Coordenador do Programa de Pós-graduação em Linguística e Língua Portuguesa da Unesp de Araraquara (SP) e membro da Comissão Permanente da Coordenadoria de Ações Afirmativas, Diversidade e Equidade (CAADI) da Unesp. Releitura. Orcid: 0000-0002-2887-3570. E-mail: matheus.schwartzmann@unesp.br

[f] Professora da Faculdade de Letras da Universidade Federal de Minas Gerais. Em 2024, realiza pós-doutorado na Universidade Federal Fluminense e na Université de Liège com financiamento Capes PrInt. Releitura final. Orcid: 0000-0003-1861-3609. E-mail: danierveLin@gmail.com

DOI: http://dx.doi.org/10.11606/issn.1982-8160.v18i3p29-54

**ABSTRACT**

Artificial intelligence today simulates the complexity of language and human actions in an increasingly satisfactory manner. In this paper I discuss artificial intelligences using semiotic tools, assuming as a theoretical standpoint É. Benveniste's theory of enunciation in post-Greimasian semiotics, notably Jacques Fontanille's concept of enunciative praxis, applied to the study of artificial intelligence. This theoretical basis will allow us to address the relation between image databases and algorithms in analyzing large image collections through computer vision, as well as user's communication modes with the Midjourney generative artificial intelligence model, focusing on machine creativity.

**Keywords:** Artificial intelligence, generative model, enunciative praxis, image generation, Midjourney

A HISTÓRIA DA INTELIGÊNCIA artificial (IA) remonta à década de 1950 e à máquina de Turing[g], 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[h].

De maneira geral, as inteligências artificiais 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[1]. Por isso, é imprescindível que a semiótica pós-estruturalista trate das linguagens artificiais e das tecnologias e práticas de automatização das ações humanas.

Este artigo está dividido em duas seções. Na primeira, tratamos da abordagem da *análise* de bancos de dados a fim de examinar o modo como a visão computacional (*computer vision*), conjuntamente a 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 (meta-imagens).

Em seguida, nos dedicaremos ao estudo da *geração automática de enunciados visuais*, isto é, às grandes coleções arquivadas em bancos de dados, usadas para produzir novos enunciados a partir de textos antigos, já sedimentados na memória coletiva, por meio de operações ou mesmo de instruções (*prompts*)[i]. Especialmente no caso da geração de novos enunciados, estudaremos algumas interações e alguns de seus produtos textuais obtidos por meio do Midjourney. Os

[g] (N.T.) 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.

[h] (N.T.) 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.

[1] 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 o livro *Turing* de Jean Lassègue (2017).

[i] (N.T.) 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.

modelos utilizados pelo Midjourney (ou mesmo pelo DALL·E, para mencionar outro exemplo) traduzem enunciados verbais (*prompts*) em enunciados visuais, ou vice-versa: produzem enunciados verbais com o objetivo de, por exemplo, descrever uma 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. Isso porque, especialmente no caso do Midjourney, a tradução não ocorre somente entre a linguagem da máquina e a linguagem humana[2]. Ela ocorre, principalmente, entre a linguagem verbal do *prompt* (o comando dado) e a linguagem visual (o produto gerado). 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.

[2] Para uma comparação de todos os modelos generativos, consulte Santaella e Kaufman (2024).

Um banco de dados 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 já realizadas*, pensando agora no mecanismo de renovação da cultura humana formalizado por Jacques Fontanille (1999) na teoria da práxis enunciativa. Utilizaremos esta teoria neste artigo, considerando que 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 (atualização/realização)*, 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.

## ANÁLISE DE SEMELHANÇAS/DISSIMILARIDADES ENTRE IMAGENS NOS BANCOS DE DADOS

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 (2020b). 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[3].

A segunda estratégia é a *aprendizagem profunda* (*deep learning*)[j], 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.

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 adotar 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[4]. 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 viemos estudando há alguns anos (Dondero, 2020) – estão relacionadas, principalmente, à análise de imagens. Obviamente, sobretudo quando o que está em jogo é a organização de grandes coleções de dados visuais (milhares de imagens), digitalizados de acordo com suas semelhanças/dissimilaridades, a máquina está sendo solicitada a realizar uma tarefa que ultrapassa a capacidade puramente humana de análise[5].

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, ao projeto do historiador da arte Aby Warburg. Em seu trabalho *Atlas Mnemosyne* (1924-1929), Warburg (2012) estudou a imagem por meio da imagem, usando como método de investigação a visualização, de modo que imagens com características comuns em termos de estilo[6] fossem dispostas próximas umas das outras em grandes painéis pretos, em função de semelhanças de composição.

[3] Ver a esse propósito o site de Lev Manovich, especialmente a seção sobre os espaços dos estilos (Manovich, 2011).

[j] (N.T.) 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.

[4] 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.

[5] 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.

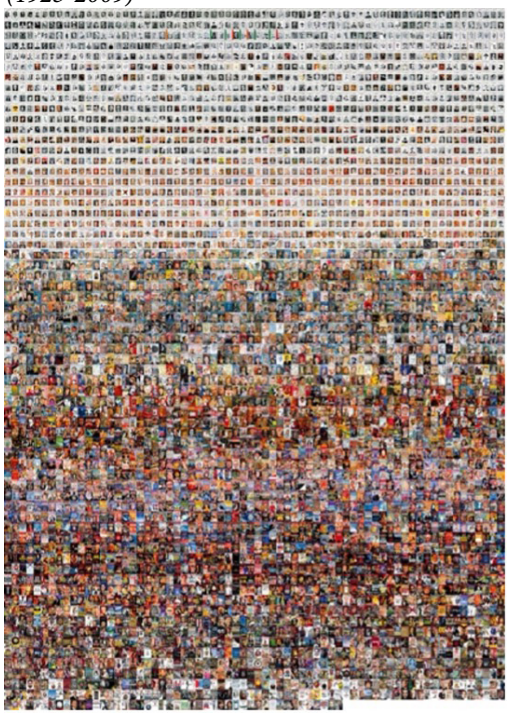
[6] Sobre o problema do estilo em Warburg, ver Pinotti (2001).

Diante das múltiplas questões que cada imagem artística coloca para o observador e para o historiador, Warburg 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 se opõe ou 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.

Há alguns anos, estudamos dois modelos de visualização (Dondero, 2017b, 2019b, 2020) que exemplificam bem as possibilidades dessa proposta. O primeiro é uma montagem clássica, com cerca de 4.500 imagens (Figura 1).

**Figura 1**

*Montagem com 4.535 capas da* Time Magazine *(1923-2009)*



*Nota*. Manovich e Douglass (2009).

Já o segundo, o mais interessante, são as visualizações que chamamos de diagrama de imagens. 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[7].

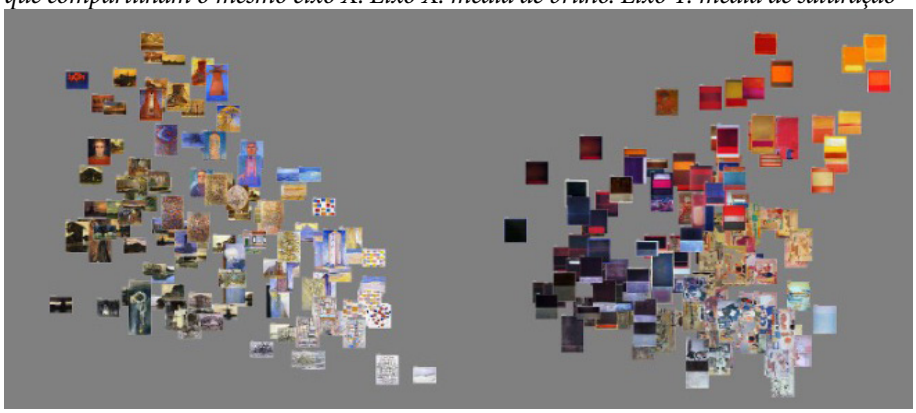
[7] Trata-se de uma estratégia que já criticamos em várias publicações (Dondero, 2017b, 2019b, 2020), nas quais explicamos que a organização de coleções por meio de metadados recai no mesmo erro pelo qual se critica Roland Barthes em relação à translinguística, ou seja, a tentativa de reduzir a imagem ao que pode ser lexicalizado.

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[8] (Figura 2).

[8] Para uma análise enunciativa desses dois tipos de visualização de imagem, consulte o terceiro capítulo de *The Language of Images* (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 proposta em *Sémiotique et littérature* (1998).

**Figura 2**

*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*



*Nota*. Manovich et al. (2011).

Dentre as propriedades plásticas, a categoria cromática é fácil de ser trabalhada pela máquina porque é uma categoria de caractere quantitativo, 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, na visão computacional, são chamadas de "recursos de baixo nível"[k]. 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 confundem, portanto, com a identificação de *formantes plásticos* ou *figurativos* (Greimas, 1984).

Em alguns trabalhos que publicamos anteriormente (Dondero, 2017a, 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*[9], deixando de se concentrar na *distribuição* dessas características dentro da imagem artística, entendida como uma totalidade[10].

Mas a despeito das críticas que podem ser feitas a esse ou àquele método estatístico, há inúmeros motivos que justificam o interesse semiótico nessas análises, dentre os quais listamos dois: essas visualizações desenvolvem uma das questões de Warburg (a das imagens e seus "vizinhos mais próximos") por

[k] (N. T.) O termo 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, dentre outros.

[9] 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.

[10] Sobre a imagem como um todo, consulte Goodman (1976), Thom (1983) e Dondero (2020).

meio de uma metodologia controlável, e também essas revelam um trabalho que pode ser entendido como estruturalista em dois sentidos:

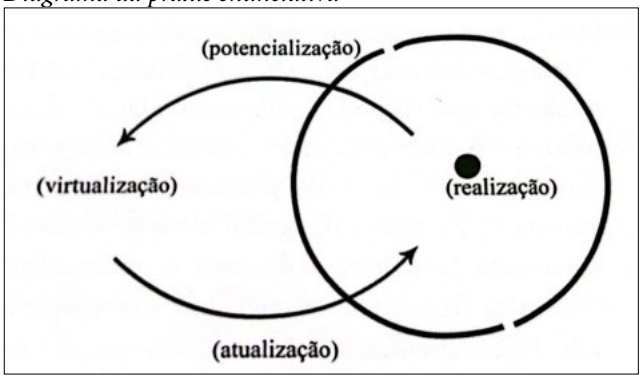
(i) essas visualizações contrastam grupos de imagens com características plásticas gradualmente semelhantes ou opostas e organizam as características de cada imagem *gradualmente*, dentro de um espaço de controle (uma perspectiva tensiva da estrutura)[11]; essas visualizações de imagens apresentam a 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);

(ii) a coleção de imagens pode ser estudada como um sistema em que a máquina foi treinada para trabalhar com diferenças e semelhanças. Como já pudemos indicar, o sistema de coleção funciona como uma enciclopédia, em outras palavras, como um sistema de "co-textos", para usar um termo de U. 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*:

[11] Ver Fontanille e Zilberberg (1998).

**Figura 3**

*Diagrama da práxis enunciativa*




*Nota*. Fontanille (1999, p. 272).

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 agora 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 *langue* saussuriana, que é composta de virtualidades, mas ele tem uma espessura conceitual muito semelhante à do *locus* da virtualização: as imagens que ele contém não são meras virtualidades pictóricas, mas imagens que foram produzidas historicamente e, de certa forma, estão copresentes no banco de dados. Elas compartilham, após a digitalização, uma substância digital comum e responsável por torná-las *comensuráveis* e disponíveis para uma 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.

## Um projeto de pesquisa sobre a genealogia dos gestos na pintura[12]

Meu projeto de pesquisa *Em direção a uma genealogia das formas visuais – Semiótica e abordagens computacionais para grandes coleções de imagens* (2022-2025, F.R.S.-FNRS) é outra maneira – mais complexa, esperamos – de dar continuidade à teorização da genealogia das formas que foi feita pelo historiador de arte Henri Focillon (1934) no livro *Vie de formes* (*Vida das formas*), e, em particular, às formas de *páthos* (o *pathosformeln*[1] 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 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). O mesmo ocorreu com a enunciação espacial – neste último caso, em especial, graças a vários trabalhos sobre perspectiva, tais como os de L. Marin (1993) e de J. Fontanille (1989). Todavia, a enunciação temporal (o antes e o 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. Há pouquíssimos estudos sobre essa questão: poderíamos mencionar o trabalho de J. Petitot (2004) sobre o *Laocoon*, um artigo do Grupo μ (1998) e outro de M. Colas-Blaise (2019), por exemplo[13].

[12] Este projeto de pesquisa, iniciado em 2022, terminará no fim de 2025. Ele está intitulado "Towards a Genealogy of Visual Forms: Semiotic and Computer-Assisted Approaches to Large Image Collections" e é financiado pelo Fonds de la Recherche Scientifique (F.R.S.-FNRS-Bélgica). Mais informações em Dondero (2022).

[1] (N.T.) 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.

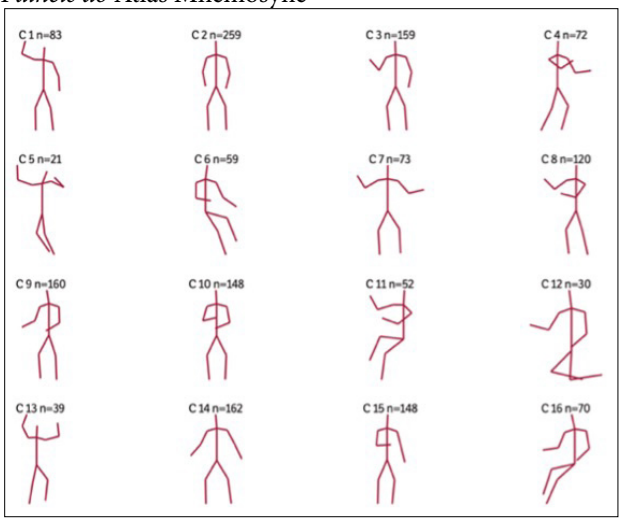
[13] Para uma discussão mais recente sobre essa questão, levando em conta os autores citados e o livro de Basso Fossali (2017), consulte Dondero (2024).

Levando em conta esse estado da arte em semiótica, nosso projeto tem o objetivo de analisar a representação de gestos corporais em um *corpus* de milhões de imagens. Por que poses e gestos corporais? Porque são o local de uma *dinâmica figurativa*, ou seja, de um *continuum,* e porque nosso desafio é o de estudar, justamente, o *movimento em imagens estáticas de maneira automática*.

Sobre esse tópico, antes de nós, Impett e Moretti (2017) formalizaram os gestos dos corpos encontrados nos painéis do *Atlas Mnemosyne*, de Warburg (Figura 4).

**Figura 4**

*Painéis do* Atlas Mnemosyne



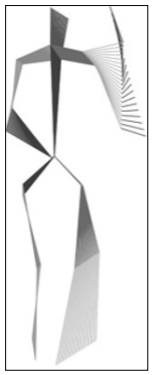

*Nota.* Impett e Moretti (2017).

Minha observação crítica em relação a esse método é que a modelagem de Impett e Moretti (2017) reduz o corpo a um esqueleto, uma figura geométrica feita de segmentos de linha, enquanto o corpo tem um volume feito de forças internas que ocupam o espaço e uma silhueta que está envolvida em cada dinâmica gestual e desempenha um papel crucial na direcionalidade dentro de uma paisagem.

Leonard Impett, em um artigo de 2020 publicado no livro *Routledge Companion to Digital Humanities and Art History*, intitulado "Analyzing Gesture in Digital Art History", tenta complexificar o modelo do corpo inserindo os parâmetros de direcionalidade e o ritmo do movimento (Figura 5). Por meio desse esquema, vemos que não apenas o esqueleto foi complexificado, mas também que Impett tenta calcular o ritmo do movimento e o deslocamento do corpo.

**Figura 5**

*Análise de componentes principais nas poses do Painel 46 do Atlas, capturando a característica morfológica mais forte do painel: a ninfa*



*Nota*. Impett (2020).

De forma semelhante, nosso projeto de pesquisa atual visa rastrear poses, gestos e outros tipos de movimento e dinâmica de forças em imagens estáticas, como pinturas e fotografias que abrangem um período que vai da pintura barroca à fotografia de moda contemporânea, extraindo poses e agrupando-as.

O que ambiciono dizer com "a dinâmica de forças em uma imagem estática"? As forças podem ser parcialmente identificadas com a direcionalidade: a direção de um olhar, de uma mão levantada, de um dedo apontado, mas também com as direções dadas por componentes da imagem que não são figurativos, mas formais/plásticos: a mudança da luminosidade em uma pintura funciona como uma espécie de seta, a mudança na saturação é capaz de produzir uma força de elevação ou de queda. A geometria de um gesto também conta: um gesto que compõe uma figura circular no plano da expressão reflete estabilidade e calma no plano do conteúdo; ao contrário, um gesto que compõe um triângulo irregular refletirá direções perturbadoras, um conflito no plano do conteúdo.

Nesse contexto, Adrien Deliège e eu tomamos 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 5 mil imagens religiosas, contendo 8.599 poses individuais. Cada pose individual é redesenhada em uma imagem separada, com coordenadas de pontos-chave normalizadas para permitir comparações significativas entre as imagens (Figura 6).

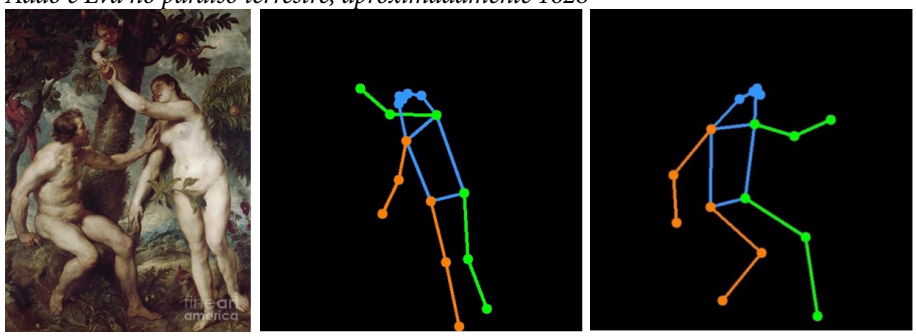
Usamos o MMPose[m] para mapeamento e o Pixplot[n] para visualização das poses. No MMPose, as poses são mapeadas a partir de 17 pontos-chave, que tentam abranger todo o corpo. Quando os 17 pontos não são contemplados, excluímos a imagem do *corpus*, pois o algoritmo não identificou um corpo inteiro.

[m] (N.T.) 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] (N.T.) 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.

**Figura 6**

*Exemplo de poses individuais extraídas de uma pintura original, Peter Paul Rubens, Adão e Eva no paraíso terrestre, aproximadamente 1628*



*Nota*. Deliège e Dondero (no prelo).

Primeiro analisamos todas essas poses individuais antes de passar para as coletivas. Definimos a distância entre duas poses como a soma das distâncias entre seus pontos-chave correspondentes. Em seguida, usamos essa métrica de distância no módulo de redução de dimensionalidade UMAP do *software* PixPlot para produzir uma visualização da organização das poses individuais (Figura 7), ou seja, as poses semelhantes estão localizadas próximas umas das outras e distantes das poses diferentes. No exemplo a seguir, é possível acompanhar essa organização.

**Figura 7**

*Visualização de 8.599 poses individuais de 5.269 pinturas religiosas contidas no banco de dados WikiArt e organizadas por similaridade de pose*[14]

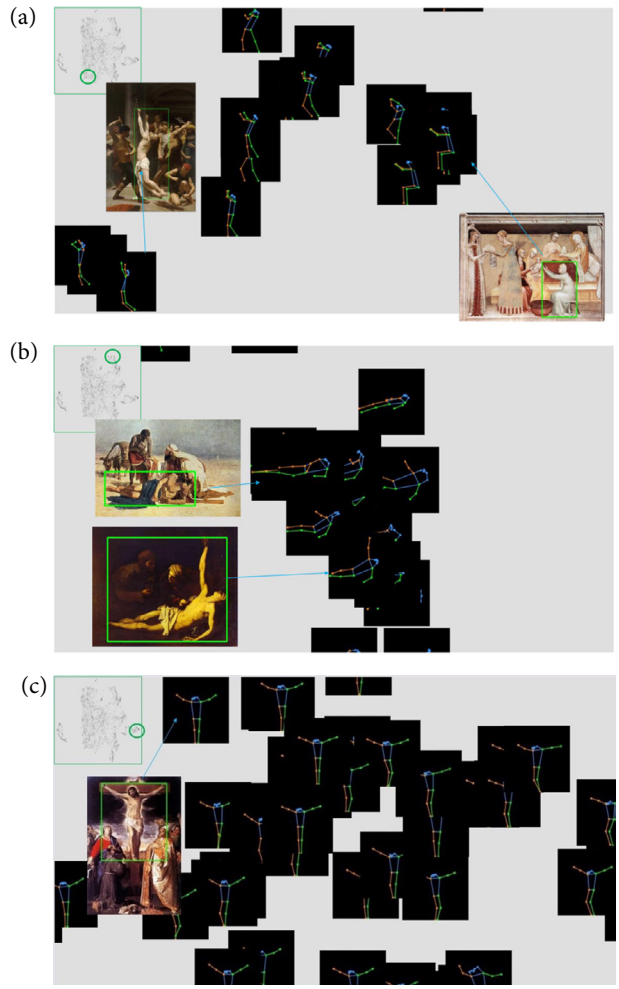
*Nota*. Deliège e Dondero (no prelo).

[14]Um aplicativo interativo da web que permite navegar por essa "nuvem de imagens" (com funcionalidades de *zoom in* e *out*). Recuperado de https://adriendeliege.z6.web.core.windows.net/outputs/WikiArt_religious_painting_solo_poses/index.html

Dentro dessa coleção, é possível acompanhar a variedade de poses de acordo com a organização estabelecida pelos algoritmos. Como exemplos, cada círculo verde no diagrama geral da coleção (canto superior esquerdo de cada subfigura da Figura 8) refere-se a um grupo de pinturas que a máquina reconhece como pertencentes a uma pose específica, que é ampliada para mostrar a validade da abordagem.

**Figura 8**

*Exemplos de poses específicas compartilhadas por determinadas pinturas, circuladas em verde (canto superior esquerdo) na visualização geral, e zoom nessas regiões delimitadas para mostrar a semelhança e as variações de poses semelhantes pertencentes a um grupo de imagens. Visualização gerada via MMPose e o Pixplot*




*Nota.* Deliège e Dondero (no prelo).

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 (circuladas em verde).

Pode-se observar que o centro da visualização tende a agrupar poses relativamente neutras, representando uma pessoa em pé, de frente para o espectador. À medida que nos afastamos do centro, as poses variam continuamente, chegando a poses completamente diferentes nos cantos da meta-imagem, como personagens deitados, sentados, caindo etc. Um grupo de representações de Jesus em sua cruz também é claramente visível, pois essa pose é comum em pinturas religiosas. Também podemos identificar um grupo de imagens em que o personagem é visto de costas, o que é completamente, e por direito, dissociado do restante das imagens (na extrema esquerda da visualização). Observemos também que um corpo deitado com a cabeça à esquerda é uma pose completamente diferente (de acordo com a métrica usada) do que se a cabeça estiver à direita da imagem. As poses opostas são representadas em partes opostas da visualização, ou seja, na parte superior e inferior nesse caso. Por fim, como em toda análise automatizada em grande escala, há alguns erros não filtrados que passaram despercebidos por essa visualização; nesse caso, como um pequeno grupo de poses com pernas cortadas nos joelhos, correspondendo a personagens que não são completamente mostrados nas imagens (na parte superior central cercada por uma vizinhança vazia na visualização).

Essas poses podem, por exemplo, ser usadas em pesquisas sobre a relação entre a expressão dessas imagens e seu conteúdo, seguindo uma *análise semissimbólica* que estabelece que uma oposição no plano de expressão está correlacionada a uma oposição no plano de conteúdo das imagens. Dois exemplos simples poderiam ser os seguintes: braços para cima *vs.* braços para baixo = oração *vs.* descanso, ou corpo em pé *vs.* corpo reclinado = pose arrogante *vs.* pose piedosa.

Ainda há várias questões a serem consideradas em poses coletivas: precisamos decidir se criaremos uma genealogia 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. Alguns dos desenvolvimentos desse trabalho podem ser encontrados em outros textos de Adrien Deliège[15].

[15] Ver Deliège (2024).

## GERAÇÃO DE NOVAS IMAGENS

Após a análise computacional de *big data*, chegamos à segunda parte deste artigo sobre a *geração automática e algorítmica de novas imagens*. Nesse caso, 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: a tradução da imagem e do texto verbal em números. 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*[o], ou, pelo menos, um modelo que entenda a linguagem natural (por exemplo, *CLIP – contrastive language-image pre-training*[j]), para transformar *prompts* (comandos) em *embeddings* (listas de números) que podem ser usados pela máquina[16]. 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.

A produção de uma série de imagens exige que o utilizador execute várias operações[17]. 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.

Também é necessário lembrar que cada imagem produzida, ou cada conjunto de imagens produzidas, é, do ponto de vista científico (o que me interessa), mais interessante como *amostras de áreas do conjunto de dados do que como imagens isoladas tout court*. Em outras palavras: as imagens produzidas pelos modelos generativos contam mais como extrações de características típicas de *uma região do conjunto de dados*, ou seja, como exemplos de padrões produzidos pelo trabalho dos algoritmos que exploram (Meyer, 2023) determinados domínios do conjunto de dados decididos pelas anotações e operacionalizados pelos *embeddings* do que como imagens estabilizadas e correspondências definitivas entre determinadas palavras e determinadas formas.

[o] (N.T.) Sistemas de inteligência artificial 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.

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

[17] Enzo D'Armenio e Adrien Deliège foram atores importantes no desenvolvimento dessas reflexões, tendo participado ativamente dos experimentos.

Mas vamos examinar agora o processo que nos leva a gerar imagens.

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: a máquina produz uma *média* de todas as pinturas dos estilos da Renascença, do Barroco, do Maneirismo e do Rococó, conforme a Figura 9.

**Figura 9**

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

(a) (c)

(b) (d)

*Nota*. Experimento realizado por Enzo D'Armenio, Adrien Deliège e M. G. Dondero (2024), via Midjourney.

Mas há duas coisas que esse processo de transformação de estilos em *médias de várias imagens* singulares não impede: a primeira é que várias médias podem produzir novas formas, conforme demonstrado por muitas competições vencidas por imagens geradas por IA[18] (o que também nos permite valorizar a parte da aleatoriedade que acompanha todas as gerações de imagens); a segunda é que, embora a máquina possa extrair e imitar vários estilos, a "mão" da máquina sempre permanece visível. Assim, podemos estudar a singularidade estilística

[18] 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", onde Jason Allen venceu o concurso de arte graças ao seu trabalho intitulado *Théâtre d'Opéra Spatial*, produzido com o Midjourney (Geoffre-Rouland, 2022). Outro exemplo de uma obra de arte produzida por inteligência artificial é *Unsupervised*, 2022, de Refik Anadol Studio. Esse projeto usa redes neurais treinadas em um banco de dados de 10 mil obras de arte da coleção do MoMA – Museum of Modern Art. Essa coleção inclui arte de 1870 a 1970, bem como obras de décadas posteriores. Sobre esse tema, indicamos o trabalho de Manovich & 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.

– o que em semiótica chamamos *opacidade enunciativa* (Marin, 1993) – da mão da máquina devido ao fato de conhecermos os estilos de referência com base nos quais a máquina trabalha[19].

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. 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/[20] etc., além, é claro, de precisar um ou mais estilos pictóricos.

Também pedimos ao Midjourney que gerasse imagens típicas de Van Gogh, por meio do *prompt*: /uma paisagem no estilo de Van Gogh/ (Figura 10). Rapidamente percebemos que seria difícil nos livrar de determinados objetos, sobretudo o sol.

**Figura 10**

*Prompt: /uma paisagem no estilo de Van Gogh/*

*Nota*. Experimento realizado por Enzo D'Armenio, Adrien Deliège e M. G. Dondero (2024), via Midjourney.

[19]Sobre a diferença entre Midjourney e DALL-E com base em seus respectivos resultados, consulte D'Armenio, Dondero, Deliège e Sarti (no prelo). Esse artigo compara vários parâmetros: a maneira como os dois modelos respondem a solicitações sobre categorias plásticas (forma, cor, topologia), sobre a relação generalidade/especificidade, sobre a temporalidade na imagem estática e assim por diante. Para uma comparação dos bancos de dados de Midjorney e DALL-E por meio de testes em iconografias pictóricas, como "Suzana e os anciãos", consulte D'Armenio et al. (2024).

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

Isso 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*/ (sem sol) ao *prompt*:

**Figura 11**

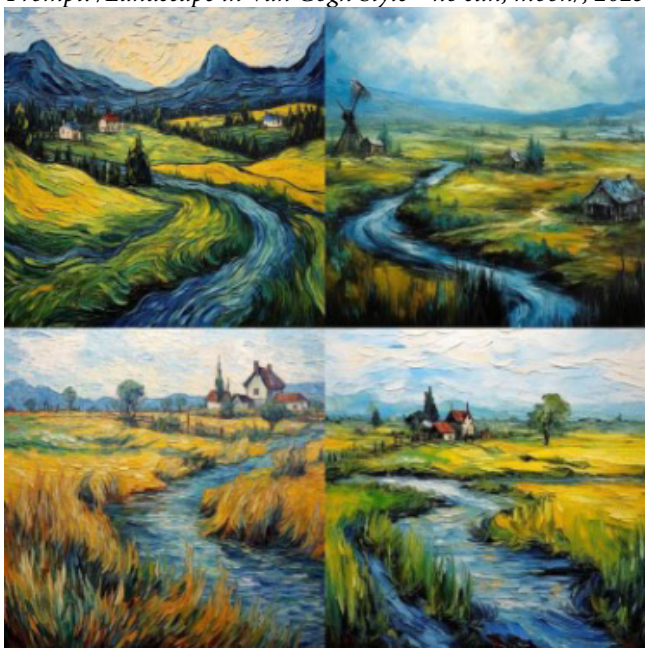
*Prompt: /uma paisagem no estilo de Van Gogh without sun without moon/*

*Nota*. Experimento realizado por Enzo D'Armenio, Adrien Deliège e M. G. Dondero (2024), via Midjourney.

Podemos ver que as imagens mantêm o sol (ou uma lua?, é difícil dizer), pois o Midjourney não foi projetado 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* (--), em outras palavras: "/*--no sun --no moon*/" (Figura 12) sem passar por mudança do *prompt*.

**Figura 12**

*Prompt: /Landscape in Van Gogh Style --no sun, moon/, 2023*



*Nota.* Experimento realizado por Enzo D'Armenio, Adrien Deliège e M. G. Dondero (2024), via Midjourney.

Consideramos esse exemplo muito significativo 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[21].

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* que se aproxima, a nosso ver, do expressionismo pictórico americano dos anos 1970.

### Misturar estilos para testar história da arte

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 relevantes que surgem com relação à composição. As imagens

[21] 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 do quê, refazer um *prompt* completo poderia modificar a imagem mais do que o desejado.

geradas por computador nos permitem entender como grandes obras de artistas do passado podem ser misturadas e, em alguns casos, apontar os estereótipos mais comuns de cada artista ou movimento artístico. Mas podemos perguntar: por que misturar estilos? Qual é o objetivo dessa mistura? Assim, podemos testar quais são os estereótipos de pintores famosos que o banco de dados aprendeu e testar combinações de estilos que revelam, pelo menos em parte, como os algoritmos trabalham na tradução entre as linguagens verbal e visual. Mas como lidamos com o fato de que a máquina pode produzir estilos ou montá-los juntos?

Como afirmou Wilf (2013) em um artigo inspirado na semiótica peirciana, escrito dez anos antes da difusão do ChatGPT e do Midjourney, intitulado "From Media Technologies That Reproduce Seconds to Media Technologies That Reproduce Thirds: A Peircean Perspective on Stylistic Fidelity and Style-Reproducing Computerized Algorithms":

> Ao contrário de um CD ou um arquivo MP3, que são tecnologias de reprodução, esses sistemas generativos não reproduzem textos específicos, ou Segundos, mas estilos, ou Terços (GENERALIDADE). Seu objeto de reprodução é o princípio da generatividade, responsável pela produção de textos específicos que são objeto de reprodução do tipo de tecnologias de mídia que tradicionalmente têm estado no centro da pesquisa antropológica linguística e semiótica. *A reprodução de estilo desses sistemas consiste tanto em sua capacidade de abstrair um estilo de um corpus de Segundos quanto de gerar Segundos ou textos novos e diferentes nesse estilo, indefinidamente* [ênfase adicionada]. (p. 186)[22]

Sem dúvida, é apenas por meio das formas já conhecidas e estabelecidas em nossa percepção cultural que é possível entender o trabalho da máquina – não apenas o grau de sua tão questionada "criatividade", mas também a maneira pela qual ela transforma as formas que conhecemos em médias. Os experimentos de Lev Manovich (publicados no Facebook em 2023) e de 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, conforme Figura 13.

Se observarmos outra produção de Manovich e Arielli (2021-2024), que mistura Brueghel e Kandinsky (Figura 14), 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 Bosch e Brueghel, tradicionalmente considerados paisagistas. Em outras palavras, se observarmos

[22] No original: "Thus, although these systems, much like a CD or an MP3 file, are technologies of reproduction, they do not reproduce specific texts, or Seconds, but styles, or Thirds. Their object of reproduction is the principle of generativity that is responsible for producing the specific texts that are the object of reproduction of the kind of media technologies that have traditionally stood at the center of linguistic and semiotic anthropological research. These systems' reproduction of style consists both in their ability to abstract a style from a corpus of Seconds and to generate new and different Seconds or texts in this style, indefinitely so". Esta tradução, dos autores.

a história da arte, há uma inversão de papéis: tradicionalmente, os pintores abstratos não são considerados artistas de paisagem, porque esse conceito não é mais relevante nesse contexto. No entanto, o trabalho da máquina usa esses pintores abstratos como estruturas que abrangem figuras inspiradas em pintores de paisagens reais, como Brueghel e Bosch.

**Figura 13**

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

Nota. Manovich e Arielli (2021-2024) via Midjourney.

**Figura 14**

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

*Nota*. Manovich & 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. 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 Monalisa 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 (Figura 15).

**Figura 15**

*Mona Lisa de L. Da Vinci + prompt: /Mona Lisa no estilo Rothko/*

*Nota*. Experimento realizado por D'Armenio, Deliège e Dondero (2024), via Midjourney 4.

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).

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 sedimentados e aceitos 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.

## CONCLUSÕES E ABERTURAS

Terminamos nossa pesquisa principalmente com perguntas. A primeira diz respeito à tradução palavra-imagem: como podemos explicar que uma inteligência generativa pode produzir uma imagem coerente em termos de composição a partir de uma solicitação verbal? A segunda, que está estreitamente relacionada à primeira, diz respeito à percepção da máquina: que tipo de percepção caracteriza a máquina geradora de imagens? Que tipo de percepção permite que ela construa uma composição bidimensional não absurda e até mesmo coerente? É uma percepção que depende dos conjuntos de dados de imagens usados no estágio de treinamento e, portanto, das imagens que relacionam a "visão" de outros, uma "visão" que é distribuída dentro do conjunto de dados? Obviamente, essa é uma percepção que calcula a média de todos os "estímulos recebidos" pelos diferentes conjuntos de dados. Como qualquer percepção, esta da máquina certamente deve ser acompanhada por uma orientação e um horizonte que, nesse contexto, é dado pelo *prompt*, que tem a tarefa de acionar os bancos de dados. Além disso, esse tipo de percepção que monta o banco de dados é manipulado por algoritmos que o direcionam. Mas que tipo de percepção é essa, já que a máquina não tem corpo nem sensações?

Embora não possamos dar uma resposta definitiva a essas perguntas muito gerais[23], que talvez possamos responder daqui a alguns anos, quando as inteligências artificiais generativas forem mais amplamente usadas e mais bem estudadas, podemos afirmar algumas coisas que nos parecem certas: os resultados

[23] Algumas respostas provisórias sobre a percepção podem ser encontradas em D'Armenio et al. (2024) e D'Armenio, Deliège e Dondero (no prelo).

não podem ser entendidos exclusivamente com base no funcionamento estatístico-computacional estrito, mas também em outros fatores mais socialmente pertinente. Mencionarei apenas dois deles aqui, no início e no final da cadeia de transformação/tradução. Esses dois fatores dizem respeito a:

(i) habilidades dos programadores em percepção humana e a ideologias envolvidas, com os programadores determinando as correspondências entre as representações numéricas aprendidas pelos modelos, na forma de listas de números chamadas de *embeddings*, e as modalidades verbais e visuais;
(ii) objetivos dos usuários desses modelos, que podem ser estéticos, artísticos, comerciais ou científicos.

Em outras palavras, e para resumir: precisa-se estudar quais operações culturais estão em ação na transição entre as correspondências palavra-imagem produzidas pelos programadores (a fase de incorporação) e a circulação de imagens geradas automaticamente (a fase de implementação diante de um público).

## REFERÊNCIAS

Beyaert-Geslin, A. (2017). *Sémiotique du portrait: De dibutade au selfie*. De Boeck Supérieur.

Colas-Blaise, M. (2019). Comment penser la narrativité dans l'image fixe? La "composition cinétique" chez Paul Klee. *Pratiques*, 181-182. https://doi.org/10.4000/pratiques.6097

D'Armenio, E., Deliège, A., & Dondero, M. G. (2024). Semiotics of Machinic Co-Enunciation About Generative Models (Midjourney and DALL·E), *Signata*, *15*. https://doi.org/10.4000/127x4

D'Armenio, E., Deliège, A., & Dondero, M. G. (no prelo). A semiotic methodology for assessing the compositional effectiveness of generative text-to-image models (Midjourney and DALL·E). In *Proceedings of the 1st workshop on critical evaluation of generative models and their impact on society, ECCV 2024*. Springer. https://orbi.uliege.be/handle/2268/321378

D'Armenio, E., Dondero, M. G., Deliège, A., & Sarti, A. (no prelo). Criteria for image generation. For a semiotic approach to Midjourney and DALL·E. *Semiotic Review*.

Deliège, A. (2024). Advances on the F.R.S.-FNRS research project "Towards a Genealogy of Visual Forms": On character poses in paintings. *Centre de Sémiotique et Rhétorique*. https://ceserh.hypotheses.org/3929

Deliège, A., & Dondero, M. G. (no prelo). The Semiotic and Computational Analysis of Represented Poses in Painting and Photography. In P. Conte, A.C. Dalmasso, M.G. Dondero & A. Pinotti (Orgs.), *Algomedia. The Image at the Time of Artificial Intelligence*. Cham; Springer.

Dondero, M. G. (2017a). Barthes entre sémiologie et sémiotique: le cas de la photographie. In J.-P. Bertrand (Org.), *Roland Barthes: Continuités* (pp. 365-393). Christian Bourgois.

Dondero, M. G. (2017b). The Semiotics of Design in Media Visualization: Mereology and Observation Strategies. *Information Design Journal*, *23*(2), 208-218. https://doi.org/10.1075/idj.23.2.09don

Dondero, M. G. (2019a). *Le travail des algorithmes. Quelques réflexions sur l'actantialité et l'énonciation* [Apresentação de trabalho]. Conferência da Associação Francesa de Semiótica. https://core.ac.uk/outputs/220155468/

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

Dondero, M. G. (2020). *The Language of Images. The Forms and the Forces*. Springer.

Dondero, M. G. (2022). P.D.R. F.N.R.S. Towards a genealogy of visual forms. *Centre de Sémiotique et Rhetórique*. https://ceserh.hypotheses.org/p-d-r-towards-a-genealogy-of-visual-forms

Dondero, M. G. (2023). Emerging Faces: The Figure-Ground Relation from Renaissance Painting to Deepfakes. In M. Leone (Org.), *The hybrid face: Paradoxes of the visage in the digital era* (pp. 74-86). Routledge.

Dondero, M. G. (2024). The Face: Between Background, Enunciative Temporality and Status. *Reti, Saperi, Linguaggi: The Italian Journal of Cognitive Sciences*. https://www.rivisteweb.it/doi/10.12832/113797

Dondero, M. G. (no prelo). Enunciación temporal en imágenes fijas. *Tópicos del seminario*.

Dondero, M. G., Rodrigues de Castro, G. H., & Schwartzmann, M. N. (2024). Inteligência artificial e enunciação: Análise de grandes coleções de imagens e geração automática via Midjourney. *Todas as Letras*, *26*(2), 1-24. https://editorarevistas.mackenzie.br/index.php/tl/article/view/17164

Eco, U. (1984). *Semiotica e filosofia del linguaggio*. Einaudi.

Focillon, H. (1934). *Vie de formes*. Presses Universitaires de France.

Fontanille, J. (1989). *Les espaces subjectifs. Introduction à la sémiotique de l'observateur*. Hachette.

Fontanille, J. (1998). *Sémiotique et littérature. Essais de méthode*. Presses Universitaires de France.

Fontanille, J. (1999). *Sémiotique du discours*. Presses Universitaires de Limoges.

Fontanille, J., & Zilberberg, C. (1998). *Tension et signification*. Mardaga.

Geoffre-Rouland, A. (2022, 2 de setembro). Midjourney: une œuvre d'art générée par l'IA remporte un concours et suscite l'indignation. *Tom's Guide*. https://www.tomsguide.fr/polemique-une-oeuvre-dart-generee-par-lia-remporte-un-concours-les-artistes-sindignent/

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett Publishing Company.

Greimas, A. J. (1984). Sémiotique figurative et sémiotique plastique. *Actes Sémiotiques Documents*, (60). https://www.unilim.fr/actes-semiotiques/3848

Groupe μ. (1998). L'effet de temporalité dans les images fixes. *Texte*, (21-22), 41-69.

Impett, L. (2020). Analyzing Gesture in Digital Art History. In K. Brown (Org.), *Routledge Companion to Digital Humanities and Art History* (pp. 386-406). Routledge.

Impett, L., & Moretti, F. (2017). Totentanz. Operationalizing Aby Warburg's Pathosformeln. *Literary Lab Pamphlet*, (16). https://litlab.stanford.edu/LiteraryLabPamphlet16.pdf

Lassègue, J. (2017). *Turing* (G. J. F. Teixeira, Trad.). Estação Liberdade.

Manovich, L. (2011, 4-6 de agosto). Style space: How to compare image sets and follow their evolution. *Manovich*. https://manovich.net/index.php/projects/style-space

Manovich, L. (2015). Data Science and Digital Art History. *International Journal for Digital Art History*, *1*(1), 3-35.

Manovich, L. (2017). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. In M. K. Schäfer, K. Vanes (Orgs.), *The Datafied Society. Studying Culture through Data*. AUP.

Manovich, L. (2020a, 22 de novembro). Computer Vision, Human Senses, and Language of Art. *AI & Society*. http://manovich.net/content/04-projects/109-computer-vision-human-senses-and-language-of-art/manovich_computer_vision.pdf

Manovich, L. (2020b). *Cultural Analytics*. MIT Press.

Manovich, L., & Arielli, M. (2021-2024). *Artificial Aesthetics: Generative AI, Art and Visual Media*. http://manovich.net/index.php/projects/artificial-aesthetics

Manovich, L., & Douglass, J. (2009). Timeline. 4535 Time Magazine Covers, 1923-2009. *Cultural Analytics Lab*. http://lab.culturalanalytics.info/2016/04/timeline-4535-time-magazine-covers-1923.html

Manovich, L., Douglass, J., & Zepel, T. (2011). How to Compare One Million Images? In D. Berry (Org.), *Understanding Digital Humanities* (pp. 249-278). Palgrave Macmillan.

Marin, L. (1993). *De la représentation*. Seuil.

Meyer, R. (2023). The New Value of the Archive: AI Image Generation and the Visual Economy of 'Style'. *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft*, *19*(1), 100-111. http://dx.doi.org/10.25969/mediarep/22314

Pinotti, A. (2001). *Il corpo dello stile: Storia dell'arte come storia dell'estetica a partire da Semper, Riegl, Wölfflin*. Mimesis.

Petitot, J. (2004). *Morphologie et esthétique*. Maisonneuve et Larose.

Santaella, L., & Kaufman, D. (2024). A Inteligência artificial generativa como quarta ferida narcísica do humano. *MATRIZes*, *18*(1), 37-53. https://doi.org/10.11606/issn.1982-8160.v18i1p37-53

Thom, R. (1983). Local et global dans l'œuvre d'art. *Le Débat*, *2*(24), 73-89.

Warburg, A. (2012). *L'Atlas Mnémosyne*. Éditions Atelier de l'écarquillé.

Wilf, E. Y. From Media Technologies That Reproduce Seconds to Media Technologies That Reproduce Thirds: A Peircean Perspective on Stylistic Fidelity and Style-Reproducing Computerized Algorithms. *Signs and Society*, *1*(2), 185-211. https://doi.org/10.1086/671751

Artigo recebido em 17 de outubro de 2024 e aprovado em 23 de outubro de 2024.