HANNES BAJOHR

L'*Ekphrasis* opératoire : l'effondrement de la distinction texte/image dans l'intelligence artificielle multimodale

Le domaine florissant des *Critical AI Studies* étend la perspective des sciences humaines au développement toujours plus rapide de ce que l'on appelle, de manière générale et quelque peu inexacte, l'« intelligence artificielle » (IA) [1]. Par son importante variété, ce champ d'étude répond à la centralité de l'IA en tant qu'apprentissage automatique stochastique dans le flux mondial de capitaux et de tendances extractives ainsi que dans la technologie de surveillance et d'exclusion raciale et économique, raison pour laquelle ces études critiques sur l'IA s'intéressent aux ramifications politiques, économiques et éthiques de ces technologies [2]. Une partie tout aussi importante des *Critical AI Studies* est consacrée à la dissection des hypothèses conceptuelles et philosophiques qui sous-tendent le développement et l'utilisation des systèmes d'apprentissage automatique, qui traitent encore souvent leurs « données » comme des représentations objectives et neutres du monde [3]. Là encore, un travail critique est nécessaire « pour détruire ce qui est supposé "naturel" et convaincre de son "artificialité" » selon les mots de Hans Blumenberg [4]. En effet, souvent, l'intelligence artificielle n'est pas considérée comme *suffisamment* artificielle. C'est dans ce creuset que les sciences humaines, armées de leur conscience critique, historique et conceptuelle, voient leur pertinence amplifiée. Comme Fabian Offert et Thao Phan l'écrivent : « les modèles d'apprentissage automatique de la génération actuelle requièrent des modes de critique (humaniste) de la génération actuelle [5]. »

1. Voir Jonathan Roberge et Michael Castelle, dir., *The Cultural Life of Machine Learning : An Incursion into Critical AI Studies* (Cham, Springer, 2021), le numéro spécial sur « Critical AI » de *American Literature*, 95, nº 2 (2023), la revue nouvellement fondée *Critical AI* 1, nºº1 (2023), et la liste de diffusion « All Models », http://allmodels.ai, consulté le 7 avril 2024.
2. Voir Kate Crawford, *Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven, Yale University Press, 2021), Wendy Hui Kyong Chun et Alex Barnett, *Discriminating Data : Correlation, Neighborhoods, and the New Politics of Recognition* (Cambridge, (Mass.) The MIT Press, 2021), Louise Amoore, *Cloud Ethics : Algorithms and the Attributes of Ourselves and Others* (Durham, Duke University Press, 2020), Safiya Umoja Noble, *Algorithms of Oppression : How Search Engines Reinforce Racism* (New York, New York University Press, 2018), et Virginia Eubanks, *Automating Inequality : How High-Tech Tools Profile, Police, and Punish the Poor* (New York, St. Martin's Press, 2017).
3. Voir Lisa Gitelman, ed. « "Raw Data" Is an Oxymoron » (Cambridge, Mass., MIT Press, 2013), Adrian Mackenzie, *Machine Learners : Archaeology of a Data Practice* (Cambridge, Mass., MIT Press, 2017) et Clemens Apprich et al., dir., *Pattern Discrimination* (Minneapolis, University of Minnesota Press, 2018).
4. Hans Blumenberg, « An Anthropological Approach to the Contemporary Significance of Rhetoric », dans *History, Metaphors, Fables : A Hans Blumenberg Reader*, dir. par Hannes Bajohr, Florian Fuchs et Joe Paul Kroll, Ithaca, Cornell University Press, 2020, p. 188.
5. Fabian Offert et Thao Phan, « A Sign That Spells : DALL-E 2, Invisual Images and The Racial Politics of Feature Space » (arXiv, 26 octobre 2022), http://arxiv.org/abs/2211.06323, p. 3.

Mais cette relation entre l'IA et les sciences humaines est réciproque : en tant qu'humanistes, nous serions négligents si nous ne mettions pas nos *propres* concepts à l'épreuve des nouveaux phénomènes que l'informatique et l'ingénierie nous proposent. Par conséquent, les pratiques humanistes doivent évoluer pour affonter les questions soulevées par la technologie de l'apprentissage automatique, et non seulement réfléchir sur et souvent contre l'IA, mais parfois aussi *avec*. Cela ne signifie pas qu'il faille abandonner la position critique, mais plutôt l'étendre aux deux côtés de l'équation et inclure les concepts humanistes comme objet d'étude et de révision potentielle, à la lumière des questions soulevées par les *Critical AI Studies*. Dans cet essai, je propose un exemple de cette « pensée avec l'IA » en éclairant d'un jour nouveau une vieille question de l'enquête humaniste : la relation entre le mot et l'image.

Dans ce qui suit, je développerai quelques intuitions autour de cette relation et je me demanderai comment elle peut évoluer dans le contexte de la transition des algorithmes classiques vers l'apprentissage automatique actuel. Je m'intéresse en particulier à ce que l'on appelle l'« IA multimodale », dont le grand modèle visuel DALL-E 2 est peut-être le plus connu. Penser *avec* l'IA revient alors à tester les ramifications théoriques de cette technologie sur l'*ekphrasis*, un concept plus traditionnel, relatif à l'interaction du mot et de l'image, que j'élargis ici pour inclure le substrat technique de cette interaction dans le numérique sous le titre d'« *ekphrasis* opérative ». En utilisant ce concept, je montre que l'IA multimodale, capable de traiter à la fois le texte et l'image comme *un seul* type de données, supprime la séparation des médias qui est au cœur de l'*ekphrasis.* Ce faisant, j'utilise l'IA comme ce que Daniel Dennett appelle une « pompe à intuition [6] », un outil qui permet de clarifier des implications conceptuelles autrement invisibles.

Dans la première partie de cet essai, je m'appuierai sur des exemples tirés de la poésie visuelle pour discuter de trois médias texte/image : 1) l'analogique, 2) le numérique « séquentiel » (informatique classique) et 3) le numérique « connexionniste » (apprentissage automatique stochastique). Je soutiens qu'avec l'avènement de l'apprentissage automatique, la division entre les médias numériques et analogiques doit être subdivisée, car l'IA fonctionne différemment des anciens paradigmes informatiques. Dans la deuxième partie, j'explique que la figure rhétorique de l'*ekphrasis* fournit un cadre pour ordonner cette nouvelle subdivision en interprétant le code comme performatif. Enfin, je tire deux conclusions : premièrement, l'opposition classique entre texte et image, sur laquelle repose le concept d'*ekphrasis*, se dissout dans l'IA multimodale ; deuxièmement, la sémantique revient néanmoins dans le numérique qui n'a jusqu'à présent été considéré que comme une question de syntaxe. Prises ensemble, ces revendications remettent en question à la fois notre lexique esthétique et notre compréhension du numérique. En tant que telles, elles soulignent

6. Daniel C. Dennett, *Intuition Pumps and Other Tools for Thinking* (New York, Norton, 2013). Selon Dennett, une pompe à intuition fonctionne en fournissant un exemple simplifié, une analogie ou une métaphore qui aide à rendre des concepts complexes plus compréhensibles et intuitifs. C'est un moyen de déclencher nos instincts et nos intuitions à propos d'une situation, afin que nous puissions comprendre plus clairement les principes sous-jacents. Dans le cas présent, la technologie concrète de l'« IA multimodale » sert de pompe à intuition pour le concept complexe d'*ekphrasis*. Cependant, contrairement au cas de Dennett, je trouve que non seulement l'exemple illustre le concept, mais qu'il est également capable de le modifier.

l'importance interdisciplinaire des *Critical AI Studies* et montrent que les sciences humaines peuvent, elles aussi, bénéficier de la réflexion sur l'IA.

TEXTE ET IMAGE DANS LE NUMÉRIQUE

Un bon moyen d'aborder la relation entre le texte et l'image est de se tourner vers la poésie visuelle, qui, par sa nature même, fait dialoguer la visualité et la textualité. La Figure 1 montre une œuvre du poète concret allemand Franz Mon. Elle est tirée de son cycle « non tot » publié en 1964 et consiste en plusieurs lignes dactylographiées ayant la forme d'un diamant, ou peut-être d'une voile. Les lignes de la moitié supérieure répètent le mot « non », celles de la moitié inférieure le mot « tot ». Les lignes se compriment progressivement vers le centre de la page, se masquant partiellement les unes les autres. La Figure 2 montre un poème visuel de l'auteur numérique allemand contemporain Jasmin Meerhoff, extrait de son recueil *They Lay* (2022). Là, des bouts de texte dactylographiés sont disposés selon un motif répétitif et ondulant qui pourrait suggérer des flammes s'élevant d'un combustible invisible. Il est difficile de déchiffrer le sens de ces lettres : il s'agit bien de lettres, mais dans leur configuration de collage, elles sont encore plus éloignées de la signification linguistique que le déjà énigmatique « non/tot » de Mon. Quoi qu'il en soit, pour le spectateur non averti, les deux œuvres convoquent un langage poétique commun qui rassemble des lettres dans des pages de constellations dont la qualité visuelle rivalise avec le sens sémantique des poèmes, voire le dépasse. Ces poèmes doivent être regardés comme des images autant (sinon plus) qu'ils sont censés être lus comme des textes. L'un à côté de l'autre, il semble que peu de choses aient changé au cours des quelque soixante années qui séparent ces deux œuvres.

Comparons cela à l'œuvre de la figure 3. Il s'agit de l'œuvre de Dave Orr qui, comme celle de Meerhoff, a été créée en 2022. Contrairement aux deux premiers poèmes, elle semble être d'une facture tout à fait différente. L'alignement centré du texte lui donne l'air d'un paradigme poétique plus traditionnel, voire naïf, antérieur à la poésie visuelle des deux autres œuvres. Pourtant, un second regard révèle que si le titre « Stiny Snity Grify » est clairement lisible, bien qu'énigmatique, les lignes ne sont pas simplement absurdes : elles ne sont pas du texte. Elles correspondent à ce que l'on appelle souvent l'écriture « asémique », c'est-à-dire une écriture qui n'utilise pas de mots, mais seulement des semblants de mots. Si, comme le dit Peter Schwenger, « les aspects visuels et musculaires de l'écriture sont généralement occultés par la primauté de la fonction communicative de l'écriture », alors un texte asémique « n'essaie pas de communiquer un message autre que sa propre nature d'écriture » [7], y compris son caractère visuel. En ce sens, le poème d'Orr pourrait lui aussi être classé comme « visuel », bien que ce soit dans une perspective différente de celle des

7. Peter Schwenger, *Asemic : The Art of Writing* (Minneapolis, University of Minnesota Press, 2019), 1.

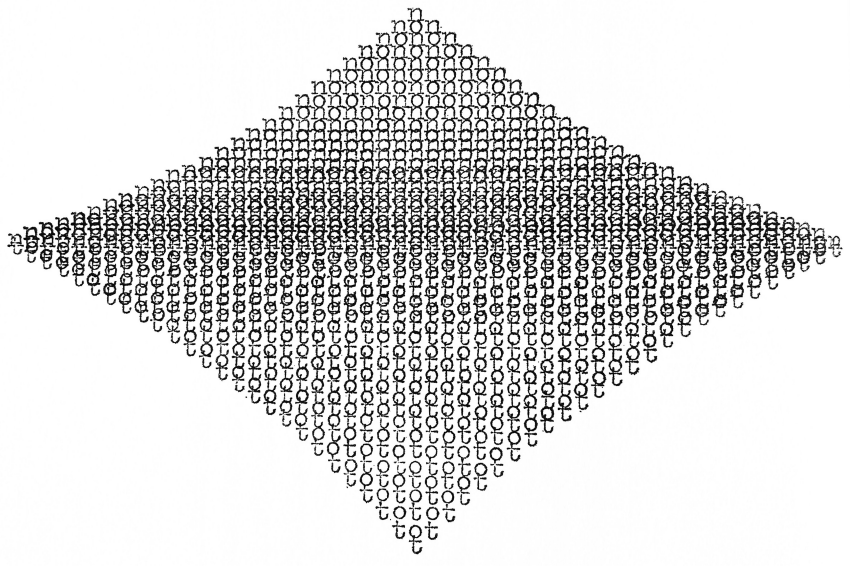

Fig.1. Franz Mon, *Non/tot* (1964).

deux autres. Au lieu de produire un poème en utilisant du texte pour créer une image, il utilise une image pour créer un poème qui ressemble à du texte.

Dans ces exemples de poésie visuelle, nous pouvons identifier des points communs entre les trois œuvres. Ce qui nous intéresse ici, en revanche, c'est ce qui les différencie à savoir leur substrat technique ou leur « spécificité médiatique » comme l'écrit Katherine Hayles [8]. En effet, les trois œuvres utilisent des technologies radicalement différentes et toutes ces technologies impliquent des relations radicalement différentes entre le texte et l'image.

Il n'est peut-être pas surprenant que les deux exemples de 2022 utilisent la technologie numérique, alors que l'œuvre de Mon a été créée en 1964 par des moyens analogiques, avec une machine à écrire mécanique Olympia Monica, pour être exact, sur laquelle il a produit une grande partie de sa poésie concrète et visuelle. La Figure 4 montre une section du manuel de la Monica avec un échantillon caractéristique de sa police de caractères. En revanche, Jasmin Meerhoff a créé son poème numériquement, en écrivant un script shell MacOS (Fig. 5). Lorsqu'il est exécuté en ligne de commande, le script demande à l'application libre Imagemagick de faire deux choses : premièrement, découper en petits morceaux un fichier image unique contenant une ligne de texte provenant d'une page scannée (lignes 12 à 20 du script) ; deuxièmement, assembler ces morceaux pour leur donner la forme qui constitue le poème (lignes 23 à 27). L'aspect ondulé est le résultat de l'utilisation d'une fonction sinusoïdale pour arranger les morceaux en spécifiant l'amplitude et la fréquence des ondes (ligne 4). Tout cela se fait automatiquement et le script de Meerhoff est disponible gratuitement en ligne [9], ce qui permet à quiconque de créer un flux potentiellement infini de poèmes visuels.

8. N. Katherine Hayles, « Print Is Flat, Code Is Deep : The Importance of Media-Specific Analysis », *Poetics Today* 25, nº 1, 2004, pp. 67-90, et Hannes Bajohr, « Algorithmic Empathy : Toward a Critique of Aesthetic AI », *Configurations* 30, nº 2, 2022, pp. 203-231.
9. Jine [i.e. Jasmin Meerhoff], « They Lay », GitLab, 24 février 2022, https://gitlab.com/nervousdata/they-lay ; consulté le 7 avril 2024.

Fig. 2. Jasmin Meerhoff, *They Lay* (2022).

L'œuvre de Dave Orr est également produite par des moyens numériques, mais de manière très différente. Elle a été créée à l'aide d'une « intelligence artificielle » ou, plus précisément, d'un algorithme complexe d'apprentissage automatique mis en œuvre sous la forme d'un réseau neuronal. Le réseau neuronal en question s'appelle DALL-E 2, un produit de la société OpenAI, mieux connue pour son modèle de génération de texte ChatGPT [10]. DALL-E 2 est un grand modèle visuel doté d'une capacité de conversion de texte en image et il n'est qu'un exemple parmi un nombre croissant d'autres générateurs, tels que

10. Aditya Ramesh et al., « Zero-Shot Text-to-Image Generation », *arXiv:2102.12092 [Cs]*, 26 février 2021, http://arxiv.org/abs/2102.12092 ; consulté le 7 avril 2024.

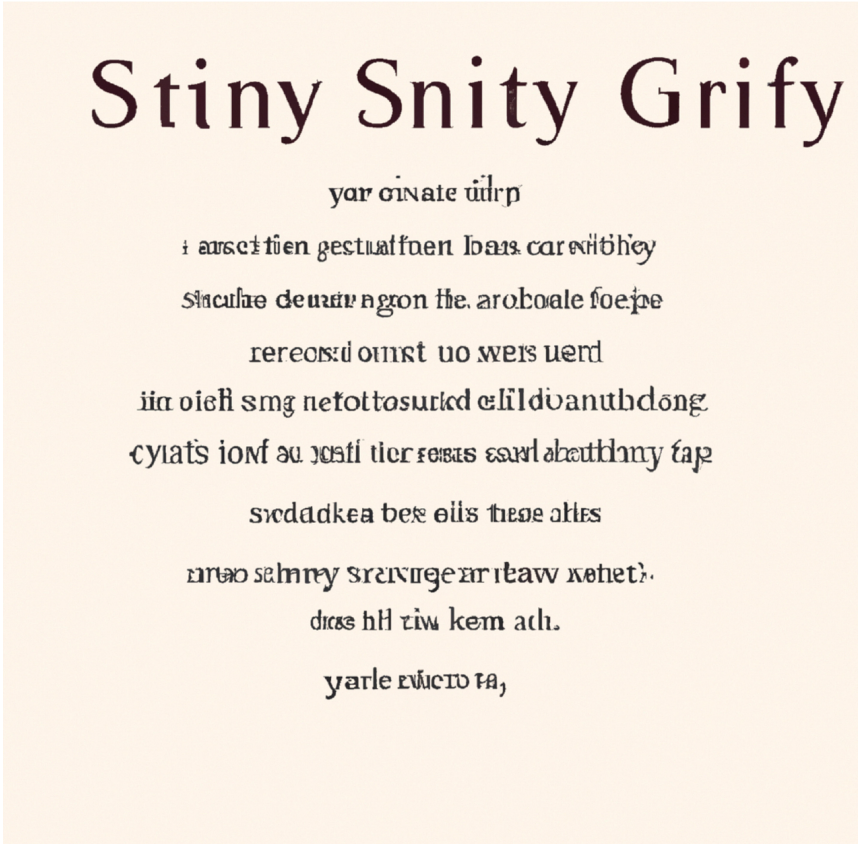

Fig. 3. Dave Orr, *Stiny Snity Grify* (2022).

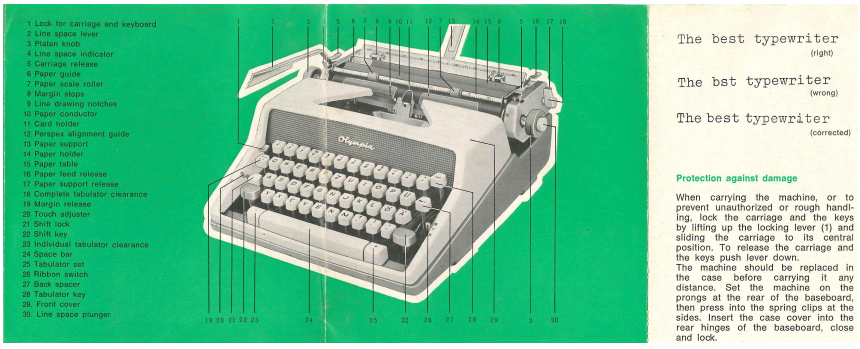

Fig. 4. Extrait du manuel de l'Olympia Monica (modèle SM7).

StabilityAI's Stable Diffusion, Google's Imagen ou Midjourney [11]. Ces systèmes prennent en *input*, en entrée, une description en langage naturel (le prompt) et génèrent une image en sortie, produisant ainsi une représentation visuelle du texte. Dans le cas de DALL-E 2, cela se fait par le biais d'une interface qui consiste en une seule zone de texte pour le prompt (Fig. 6). Pour le poème d'Orr, le prompt était « un poème sur la singularité écrit dans une police à empattement [12] ».

11. Pour en savoir plus sur les T2I, voir le numéro spécial « Generative Imagery : Towards a "New Paradigm" of Machine Learning-Based Image Production » de *IMAGE. The Interdisciplinary Journal of Image Sciences*, 31, n° 1, 2023.
12. Dave Orr, « Playing with DALL-E 2 », Lesswrong, https://www.lesswrong.com/posts/r99tazGiLgzqFX7ka/playing-with-dall-e-2 ; consulté le 7 avril 2024.

```
#!/bin/bash
# They lay (oscillating cuts) – “micro”

read -p "Enter value for AM (between 1 and 100). Enter value for FM (1 to 30000) " am fm
echo "AM is $am and FM is $fm"

read -p "Enter filename " fl
echo "File is $fl"

ct=0
until [ $ct -gt 599 ] # sets how often a cut will be made (+1) and how many snippets will be produced
do
  ((ct+=1)) # a counter. starting at 1, increasing by 1
  ctt=`printf %03d $ct` # prints 3 digits numbers, important for naming the files
  zw=`awk -v x="$ct" -v f=$fm -v a=$am 'BEGIN {wz=sin(5*(3+x*a))*sin(2*3.1416*(3+(x/f)))+0.9999; printf wz }'` # defines a sine function. shift 0.9999 above zero on y-axis
  sw=`awk -v x="$ct" 'BEGIN {wv=(sin(x*4)+5)*0.28; printf wv }'` # defines another sine function for slightly changing the width of snippets
  mkdir -p cuts pieces # creates directories for the cut images and the new images
  echo " Cutting …"
  magick $fl +profile "icc" -gravity SouthWest -crop "%[fx:(w/20)*$sw]"x"%[fx:h]"+"%[fx:(w/2.75)*$zw]"+0 +repage cuts/$ctt.png # cuts the image. the width of the snippets is set as a fraction of the width of the input image. the position on the x-axis defines where the cut is made. it is calculated with the values of the variable ‘zw’
done

while :
do
  echo ' Assembling ! '
  montage cuts/*.png -tile 20x -background white -geometry +0+0 -units PixelsPerInch -density 300 pieces/thl_am"$am"_fm"$fm".png # composes a new image. with -tile the number of snippets in a row is defined. it should be a (integer) divisor of the number of snippets (line 7) to avoid gaps at the bottom of the new image
  break
done
```

Fig. 5. Le script shell pour *They Lay* de Jasmin Meerhoff.

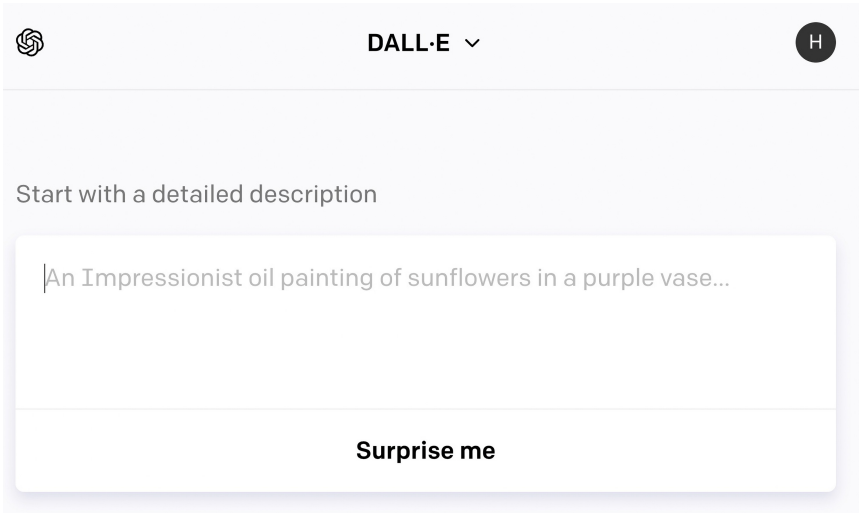

Fig. 6. L’interface de DALL-E 2 avec une zone de texte pour saisir l’invite.

Il convient de noter que DALL-E 2 ne génère généralement pas de textes. Ce « poème » est apparu lorsque M. Orr examinait le modèle et il a été publié dans le cadre d’un article de blog sur le système. Il semble qu’il n’ait jamais été destiné à être publié en tant qu’œuvre littéraire et le fait qu’Orr soit chef de produit pour Google Assistant et ne se considère pas, à ma connaissance, comme un poète, confirme cette impression. En effet, DALL-E 2 est réputé

Fig. 7. Une image générée par DALL-E 2 pour l'invite « Un astronaute à cheval en style photoréaliste ».

pour ne pas produire de texte et ses images de mains mutilées ainsi que l'écriture brouillonne sont (ou étaient jusqu'à récemment) le signe le plus évident qu'une image est générée par l'intelligence artificielle [13]. DALL-E 2 est généralement censé produire des images photoréalistes ou stylisées qui ont toutes en commun d'être le résultat d'un texte d'entrée. La Figure 7 montre un exemple plus classique tiré du site web du développeur. Le prompt « Un astronaute à cheval en style photoréaliste » donne lieu à une image de ce type. Comme il n'y a rien d'autre que le prompt textuel pour orienter la génération d'images, une véritable « promptologie » s'est établie depuis la popularisation des grands modèles visuels. En peaufinant le texte d'entrée, en ajoutant des descriptions de style ou d'atmosphère, il est possible d'orienter le résultat dans une direction ou une autre. Outre le *livre de prompteurs de DALL-E 2* [14], il existe même un site Internet sur lequel des prompts particulièrement utiles sont vendus à des prix modiques [15].

Les trois œuvres discutées incarnent chacune des paradigmes poétiques et technologiques différents, que l'on peut classer de plusieurs manières : 1) comme un type d'interaction texte/image dans le genre plus large de la

13. Sur la question des mains mutilées, voir Amanda Wasielewski, « Midjourney Can't Count« », *IMAGE*, 37, n° 1, 2023, pp. 71–82 (https://doi.org/10.1453/1614-0885-1-2023-15454 ; consulté le 7 avril 2024). Pour l'incapacité à produire du texte, voir Eliza Strickland, « DALL-E 2's Failures Are the Most Interesting Thing About It », *IEEE Spectrum*, 14 juillet 2022 (https://spectrum.ieee.org/openai-dall-e-2 ; consulté le 7 avril 2024) Toutefois, cela peut dépendre de la taille des paramètres : le modèle Parti de Google semble pouvoir produire du texte avec un nombre de paramètres supérieur à 20 milliards : « Parti : Pathways Autoregressive Text-to-Image Model », Google Research, consulté le 12 juillet 2023, https://sites.research.google/parti/ ; consulté le 7 avril 2024
14. Dall-ery gall-ery, « The DALL-E 2 Prompt Book, v1.02 », *Dall-ery gall-ery : Ressources for Creative DALL-E Users*, 2022, https://dallery.gallery/wp-content/uploads/2022/07/The-DALL%C2%B7E-2-prompt-book-v1.02.pdf ; consulté le 7 avril 2024
15. PromptBase, https://promptbase.com, consulté le 12 juillet 2023.

poésie visuelle et 2) comme le résultat d'une technologie analogique (Mon) ou numérique (Meerhoff, Orr). Cependant, il est possible de diviser la technologie numérique en deux sous-catégories : 3) les algorithmes classiques et l'IA moderne, que je discuterai dans un instant sous la rubrique des paradigmes respectivement séquentiel et connexionniste. Il est d'ores et déjà possible de constater que le domaine numérique n'est pas un monolithe, mais plutôt un paysage de sous-domaines variés.

Les trois classifications relient le texte à des images ou à des structures semblables aux images, mais elles le font de manière distincte. La poésie visuelle le fait par sa nature même : elle crée des images par l'agencement du texte. Cependant, seules les deux œuvres numériques le font au niveau du substrat technique. Là, le texte et l'image-texte qui en résulte ne sont pas simplement dans une relation mimétique, mais dans une relation causale provoquée par un langage de code purement syntaxique. C'est ce que j'appelle l'*ekphrasis opérative*. Mais ce n'est que dans le dernier cas, le modèle de l'IA, que l'on trouve un élément sémantique qui menace de dissoudre complètement la distinction entre le texte et l'image. Le reste de cet essai sera consacré à l'analyse de ces distinctions. Elles ont des implications significatives sur la manière dont nous interprétons et comprenons ces œuvres et leurs différences et c'est un exemple de ce que pourrait être une esthétique de l'IA qui prendrait le substrat technique au sérieux. Pour comprendre pourquoi, il me faut élucider dans une certaine mesure le fonctionnement de ces technologies.

PARADIGMES SÉQUENTIELS ET CONNEXIONNISTES DE L'IA

Pour la subdivision de la technologie numérique, j'ai proposé les termes « séquentiel » et « connexionniste » [16]. Le paradigme séquentiel désigne le style dominant de fonctionnement des ordinateurs depuis l'invention (conceptuelle) de la machine universelle par Alan Turing en 1936 et, après la construction d'anciens modèles, la mise en œuvre (effective) par John von Neumann du concept de « programme stocké » dans l'architecture EDVAC en 1945 (construite en 1949) [17] qui, dans l'ensemble, est encore utilisée aujourd'hui. Il se caractérise par l'algorithme classique, défini dans un langage de programmation d'étapes exécutées de manière séquentielle. Le *cut-up script* de Meerhoff appartient à cette catégorie, tout comme la plupart des programmes d'un ordinateur classique. Par exemple, la commande « read -p » à la ligne 4 demande à l'utilisateur d'entrer des données qui seront stockées dans les variables « am » et « fm », qui désigneront plus tard l'amplitude et la fréquence des ondes du poème. Il est important de noter que ces lignes sont exécutées l'une après l'autre et de manière déterministe. Chaque fois qu'il est exécuté, le programme passe

16. Je tire cette distinction conceptuelle d'une publication influente qui a remis les réseaux neuronaux à la mode sous la rubrique du « connexionnisme » : David E. Rumelhard, James McClelland et Geoffrey Hinton, « The Appeal of Parallel Distributed Processing », dans David E. Rumelhard, James McClelland et PDP Working Group, (eds.), *Parallel Distributed Processing. Explanations in the Microstructure of Cognition, Vol. 1, Foundations* (Cambridge, MA, MIT Press, 1986), p. 43. Le terme « séquentiel » pour l'algorithme classique provient du même livre : David E. Rumelhart et James L. McClelland, « PDP Models and General Issues in Cognitive Science », *ibidem*, p. 116.
17. Voir Thomas Haigh et Paul E. Ceruzzi, *A New History of Modern Computing*, Cambridge (Mass.), MIT Press, 2021.

par les mêmes commandes prévisibles. Comme il est possible d'inspecter l'algorithme en lisant les règles explicitement énoncées, ce paradigme présente, en principe, un degré élevé de transparence pour les lecteurs humains.

Le paradigme séquentiel diffère grandement du mode de fonctionnement numérique plus récent que j'appelle connexionniste, qui est ce que l'on entend généralement par IA aujourd'hui c'est-à-dire, l'apprentissage en profondeur par le biais de réseaux de neurones artificiels. S'inspirant librement de la manière dont les neurones individuels du cerveau établissent des chemins pour exécuter des fonctions de plus haut niveau, les réseaux neuronaux profonds actuels sont eux aussi constitués de connexions de « neurones » et de « synapses », calculant et mettant à jour les valeurs d'activation entre ces neurones. Il est important de noter, cependant, qu'il s'agit d'une affaire hautement idéalisée et qu'il ne faut pas la confondre avec la structure réelle du cerveau. Au fur et à mesure que le modèle est formé à l'aide de données dans le cadre d'un processus d'apprentissage itératif, une valeur, ou « poids », est attribuée à chaque « neurone » du réseau neuronal computationnel. Par un processus connu sous le nom de « propagation vers l'avant », le réseau neuronal utilise les données initiales pour calculer une sortie. L'écart entre cet *output*, cette sortie et la sortie correcte souhaitée ou connue est ensuite mesuré à l'aide d'une fonction de perte. Ensuite, un algorithme d'optimisation, généralement une variante de la « descente de gradient », est utilisé pour ajuster les poids et les biais du réseau afin de minimiser cette perte par un processus connu sous le nom de « rétropropagation ». Le but ultime de l'entraînement d'un réseau neuronal est d'approximer une fonction qui peut généraliser correctement à partir des données de formation vers des données auxquelles elle n'a pas encore accès. En d'autres termes, le réseau est chargé d'identifier des modèles ou des structures sous-jacents dans les données d'apprentissage qui lui permettent de produire des prédictions ou des classifications précises lorsqu'il est confronté à de nouvelles données similaires [18].

Prenons un exemple pratique impliquant la génération d'images. Un réseau neuronal peut traiter un ensemble de données suffisamment important de visages humains pour en apprendre les modèles, les structures et les variations inhérents. Ces caractéristiques apprises peuvent ensuite être appliquées pour générer des images de visages entièrement nouvelles qui, bien qu'elles soient totalement inédites, ressembleront de manière frappante à des visages humains réels. En raison de la nature statistique de l'IA, ces visages ne sont ni des collages de parties de visages, ni de simples composites linéaires de tous les visages connus du modèle. Au contraire, et de manière métaphorique, le réseau apprend la *nature du visage*, la *Gestalt* des visages et est capable de la recréer d'une manière qui ne répète pas les entrées individuelles [19]. C'est le principe du célèbre site web thispersondoesnotexist.com, qui présente un portrait

18. Pour une introduction non technique, voir John D. Kelleher, *Deep Learning*, Cambridge (Mass.), MIT Press, 2019. Pour une discussion plus technique, voir Ian Goodfellow, Yoshua Bengio et Aaron Courville (dir.), *Deep Learning*, Cambridge (Mass.), MIT Press, 2016.
19. Voir Hannes Bajohr, « The Gestalt of AI. Beyond the Holism-Atomism Divide », *Interface Critique*, nº 3, 2021.

complètement nouveau et unique, mais artificiel, d'un visage à chaque fois qu'il est actualisé.

Le modèle d'IA résultant de ce processus de formation met en œuvre des fonctions non linéaires complexes. Cependant, un réseau neuronal ne peut pas être traduit en un algorithme déterministe et exact, car le modèle décrit simplement les forces de connexion entre les « neurones » dans ce que l'on appelle le modèle de poids. Bien que les réseaux neuronaux soient toujours mis en œuvre dans une machine numérique de von Neumann, et non par exemple dans un ordinateur analogique ou quantique, et qu'ils soient donc toujours numériques, ils suivent néanmoins un cadre conceptuel radicalement différent de celui du modèle séquentiel. En effet, contrairement au paradigme séquentiel, dont la logique est présentée étape par étape, le paradigme connexionniste suit une logique stochastique plutôt que purement déterministe. Les connaissances acquises sont intégrées dans la structure du réseau et dans ses poids, qui représentent la force des connexions entre les neurones artificiels. Par conséquent, bien qu'il soit techniquement possible de « lire » les valeurs des poids dans un réseau neuronal entraîné, ces chiffres ne se traduisent pas par une séquence d'instructions ou d'étapes compréhensibles comme le fait un code traditionnel dans un langage de programmation [20].

De toute évidence, il s'agit de deux modèles de calcul très différents que nous appelons néanmoins « numériques ». Le travail de Meerhoff, produit par un algorithme classique, était un exemple du paradigme séquentiel, tandis que le « poème » d'Orr, produit par un réseau neuronal, était un exemple du paradigme connexionniste. Cette introduction technique est importante car cette distinction imbriquée (celle entre l'analogique et le numérique et celle au sein du numérique) entre le paradigme séquentiel et le paradigme connexionniste génère des relations différentes entre le texte et l'image. En effet, les deux formes numériques, mais pas la forme analogique, se caractérisent par ce que j'appelle « *l'ekphrasis* opérative ».

LES NOTIONS DE REPRÉSENTATION ET DE PERFORMANCE DE L'*EKPHRASIS*

Le concept d'*ekphrasis* est l'un des termes les plus discutés dans la théorie visuelle, la critique littéraire et les classiques pour décrire la relation entre le texte et l'image. Comme l'a noté Ruth Webb, il est devenu un genre théorique à part entière, évoquant

> un réseau de questions et d'intérêts interdépendants, depuis la recherche positiviste de monuments perdus décrits dans les *ekphrasis* antiques et médiévales jusqu'à la fascination poststructuraliste pour un fragment textuel qui se déclare être un pur artifice, la représentation de la représentation [21].

20. Voir Fabian Offert, « Can We Read Neural Networks ? Epistemic Implications of Two Historical Computer Science Papers », *American Literature*, 95, nº 2, 2023, pp. 423-428 ; https://doi.org/10.1215/00029831-10575218 ; consulté le 7 avril 2024.
21. Ruth Webb, « *Ekphrasis* Ancient and Modern : The Invention of a Genre », *Word & Image*, 15, nº 1, janvier 1999, p. 7.

Mais il a été souligné à maintes reprises que, dans son sens premier, l'*ekphrasis* était une catégorie beaucoup plus large et signifiait un dispositif rhétorique pour générer des descriptions vives et sensorielles dans l'art oratoire. Ce terme, issu de l'éducation du rhéteur, était utilisé pour décrire l'acte de faire apparaître clairement quelque chose dans l'esprit de l'auditoire, de le transformer, selon les mots de Nicolas de Myre au IIe siècle après J.-C., d'auditeur en spectateur [22].

Cette première signification implique déjà une anthropologie des médias dans laquelle les sens auditif et visuel deviennent fonctionnellement interchangeables. L'usage ancien ne s'attachait pas particulièrement à la description d'œuvres d'art visuelles, comme le souligne Webb [23]. Ce n'est que plus tard, au XIXe siècle et surtout au XXe siècle, que le terme *ekphrasis* a été restreint aux représentations *littéraires* d'une œuvre d'art réelle ou, dans le cas de ce que John Hollander a appelé plus tard « *ekphrasis* notionnelle », d'une œuvre d'art imaginaire [24]. Néanmoins, les deux interprétations persistent sous différentes formes jusqu'à aujourd'hui, de sorte que, au cours des cinq dernières décennies, les définitions de l'*ekphrasis* ont varié de « toute description de quelque chose de visuel [25] » à, plus spécifiquement, « la description poétique d'une œuvre d'art picturale ou sculpturale [26] ».

Si l'on s'en tient à la définition la plus large et je pense, la plus philosophiquement générative, il est logique que James Heffernan ait mis l'accent sur la qualité *représentative* de l'*ekphrasis* en la définissant comme « la représentation verbale d'une représentation visuelle [27] ». Tamar Yacobi a souligné ce point de vue en suggérant que l'*ekphrasis* est une « représentation au second degré » en spécifiant que la représentation est une répétition dans un mode différent : « Ce qui était à l'origine une image autonome du monde devient dans le transfert ekphrastique une image d'une image, une partie d'un nouveau tout, une *insertion* visuelle dans un *cadre* verbal [28]. » Nous pouvons d'ores et déjà constater que cette caractérisation, ainsi que la rhétorique de l'insertion et du cadre, mettent en évidence la tension qui se trouve au cœur de l'*ekphrasis* : Le concept articule soit une équivalence, soit une compétition entre le langage et l'image. L'imitation de l'un par l'autre est soit une entreprise réussie, soit la recette d'une déception. Ainsi, comme l'a noté W. J. T. Mitchell, l'*ekphrasis* peut faire partie d'une ontologie de l'interaction texte/image qui suscite l'espoir ou la crainte. Il s'agit soit d'un potentiel de transformation presque utopique du visuel au verbal et vice-versa, comme le voulaient les anciens, soit d'une impossibilité flagrante, qui doit donc être interdite sur le plan esthétique : rendre le visuel absolument verbal ne peut jamais se produire en réalité et cela ne doit pas se produire [29]. Le *Laocoon* de Lessing [30] défend l'incompatibilité entre la structure temporelle du langage (idéale pour représenter l'action) et la composition spatiale de la peinture (mieux adaptée à la représentation d'objets)

22. Voir Ulrich Pfisterer, « Ekphrasis », dans *Metzler Lexikon Kunstwissenschaft : Ideen, Methoden, Begriffe*, Ulrich Pfisterer (dir.), Stuttgart, J. B. Metzler, 2019, p. 99 ; voir aussi Webb, « *Ekphrasis* Ancient and Modern : The Invention of a Genre », art. cit.
23. Webb, *ibidem*, p. 8.
24. John Hollander, *The Gazer's Spirit : Poems Speaking to Silent Works of Art*, Chicago, The University of Chicago Press, 1995, p. 7.
25. James A. W. Heffernan, « *Ekphrasis* : Theory », dans *Handbook of Intermediality : Literature, Image, Sound, Music*, Gabriele Rippl (dir.), Berlin, de Gruyter, 2015, p. 35.
26. Leo Spitzer, « The "Ode on a Grecian Urn", or, Content vs. Metagrammar », dans *Essays on English and American Literature*, dirigé par Anna Hatcher, Princeton, Princeton University Press, 1962, p. 72.
27. James A. W. Heffernan, Museum of Words: The Poetics of Ekphrasis from Homer to Ashbery, Chicago, University of Chicago Press, 1993, 2004, p. 3.
28. Tamar Yacobi, « Ekphrastic Double Exposure and the Museum Book of Poetry », *Poetics Today*, 34, nº 1-2, 2013, pp. 1, 3.
29. W. J. T. Mitchell, *Picture Theory : Essays on Verbal and Visual Representation*, Chicago, University of Chicago Press, 1994, pp. 152-160.
30. Gotthold Ephraim Lessing, *Laocoon : An Essay upon the Limits of Painting and Poetry*, trad. Ellen Frothingham, Mineola, Dover, 2005, chap. 15 et 16.

et constitue, selon Mitchell, « l'expression classique de la peur ekphrastique [31] ».

Cette analyse des caractéristiques immanentes des médias impliqués dans la métaphore de la « peinture avec des mots » (Horace) implique une critique de la représentation mimétique en tant que pivot de l'*ekphrasis*. Elle a été reprise au cours de la dernière décennie et l'accent mis sur la représentation a été remplacé par un accent mis sur la *performance* : qu'est-ce que l'*ekphrasis fait*, sans dire que ce faire doit être imitatif ? Renate Brosch a proposé une nouvelle définition : « L'*ekphrasis* est une réponse littéraire à une image visuelle… qui met l'accent sur la performance plutôt que sur le mimétisme [32]. » Cette interprétation performative de l'*ekphrasis* présente plusieurs avantages, le principal étant qu'en passant outre sa dimension mimétique, on peut suspendre la décision quant à son interprétation pleine d'espoir ou de crainte. Au lieu de la comprendre comme une équivalence ou une compétition entre les formes d'art, comme une relation de représentation réussie ou non, elle les place simplement dans une relation consécutive et causale.

J'aimerais étendre cette notion en mobilisant la définition performative de l'*ekphrasis* pour les médias numériques. Avec un accent différent, Brosch transpose également l'*ekphrasis* au numérique, en soutenant qu'elle devient importante dans une écologie des médias numériques qui est inondée d'images tout en se noyant dans le texte, contredisant ainsi les prédictions apocalyptiques qui annonçaient la disparition de la lecture. Cependant, je modifierais son utilisation du mot « performatif » pour parler de l'*ekphrasis* dans sa spécificité médiatique au sein du numérique. Au lieu de « réponses » littéraires, qui ne sont pas elles-mêmes des événements numériques, je veux comprendre la performance de l'*ekphrasis* comme une *opération informatique qui met en corrélation le texte et l'image.*

EKPHRASIS OPÉRATOIRE

En gardant à l'esprit la notion performative d'*ekphrasis*, revenons aux trois poèmes visuels, à la division entre analogique et numérique et à la subdivision entre séquentiel et connexionniste. Ces trois œuvres incarnent des manières spécifiques d'utiliser le langage pour créer une image. En ce sens, elles sont toutes ekphrastiques dans leur effet cumulatif : produire des constellations visuelles à travers le texte. Cependant, ce n'est pas encore ce que j'appelle l'*ekphrasis* opérationnelle. Ce n'est que dans les œuvres numériques, et non dans les œuvres analogiques, que le texte peut réellement faire naître une image de manière active et causale.

31. Mitchell, *Picture Theory : Essays on Verbal and Visual Representation, op. cit., p.* 154.
32. Renate Brosch, « *Ekphrasis* in the Digital Age : Responses to Image », *Poetics Today*, 39, n° 2, 2018, p. 227.

Dans l'analogique, c'est à dire dans le poème de Mon sur la machine à écrire, le texte peut bien sûr « produire » une image. Cependant, cette production n'est pas performative au niveau opérationnel, mais plutôt un effet perceptif après coup d'un arrangement manuel. Dans « non/tot », ce sont les actions corporelles de l'écrivain, les mouvements de ses mains sur la machine à écrire, la force qu'il exerce sur les touches, qui conduisent à ce que nous sommes obligés de décrire comme l'« image » du texte. Cette structure visuelle est le résultat d'un travail, c'est-à-dire d'une chaîne causale de forces mécaniques qui ne sont pas elles-mêmes textuelles. Il n'y a ici qu'un texte, celui de la page, qui n'accomplit rien à proprement parler.

Il en va différemment dans les œuvres numériques. Dans l'œuvre de Meerhoff, il y a désormais deux textes : celui sur la page et celui qui produit effectivement ce texte, le script. Il s'agit d'une performance textuelle au sens *computationnel* du terme : une opération que le premier texte effectue pour produire effectivement le second texte. Il ne s'agit pas d'un travail mécanique, comme dans Mon, mais de la manipulation d'informations qui sont elles-mêmes de nature textuelle. Cette idée n'est pas nouvelle, bien sûr, et des chercheurs comme Espen Aarseth ont construit des théories entières autour de cette dualité du texte [33], tandis que Katherine Hayles a soutenu que « le texte électronique est plus processuel que l'imprimé, il est performatif de par sa nature même [34] ». C'est précisément cette performativité de l'interaction entre le premier texte (le code) et le second texte (la constellation-image finale) que j'appelle l'*ekphrasis* opérative. Il s'agit de comprendre l'*ekphrasis* non pas comme une représentation mais comme une performance ; non pas comme l'imitation d'une image par un texte, mais comme un texte qui produit effectivement une image. En tant que telle, il s'agit véritablement de « mots qui peignent une image », mais en tant qu'opération de manipulation d'informations symboliques plutôt qu'en tant que représentation figurative [35].

Deux remarques s'imposent ici pour répondre aux objections possibles à cette notion d'*ekphrasis* opératoire. Premièrement, il est facile de noter que ce qui est « peint » ici n'est pas en fait une image *textuelle.* L'œuvre de Meerhoff peut utiliser du texte (le code) pour créer une image composée de texte (l'œuvre), mais, techniquement, le résultat est un fichier image et non un fichier texte. Son contenu, une fois affiché sur un écran d'ordinateur, n'est enregistré comme du texte que pour les humains, mais pas pour les machines. Il s'agit d'une image bitmap, une grille de pixels avec différentes valeurs de couleur, et en tant que telle, elle est lisible par l'homme, mais pas par la machine. Sans un processus de reconnaissance optique des caractères, l'ordinateur ne l'enregistrerait même pas comme du texte.

Pour répondre à cette objection, il convient de noter que l'image est, elle aussi, constituée textuellement à un niveau inférieur : les fichiers d'images sont

33. Espen J. Aarseth, *Cybertext : Perspectives on Ergodic Literature*, Baltimore, The Johns Hopkins University Press, 1997.
34. N. Katherine Hayles, *My Mother Was a Computer*, Chicago, The University of Chicago Press, 2005, pp. 101, 50.
35. L'adjectif « opérative » doit donc être compris littéralement comme « ayant le caractère d'une opération« ». Il ne doit pas être confondu avec l'*operatives Bild* de Harun Farocki, parfois traduit par « opératif » ou « image opérationnelle », par lequel il désigne les images utilisées dans la surveillance et la guerre qui ne nécessitent pas de médiation linguistique parce qu'elles agissent comme des capteurs plutôt que comme des représentations, Harun Farocki, « Phantom Images », *Public*, 29, 2004, https://public.journals.yorku.ca/index.php/public/article/view/30354 ; consulté le 7 avril 2024. Alors que Jussi Parikka, reprenant l'idée de Farocki, souligne la performativité des images, mon concept rend le texte performatif dans la mesure où il produit une image, voir Jussi Parikka, *Operational Images : From the Visual to the Invisual*, Minneapolis, University of Minnesota Press, 2023.

en effet codés de manière alphanumérique. Ce n'est qu'en traduisant ce texte dans la matrice de pixels d'un écran au moyen d'un codec qu'il devient effectivement une image [36]. Cet argument a été avancé pour décrire l'*ekphrasis* numérique dès 1996, lorsque le théoricien des médias Jay David Bolter a déclaré que si la tradition de l'interaction texte/image était fondée à la fois sur la supériorité du mot sur l'image et sur une métaphysique de la présence qui espérait atteindre la chose elle-même par une description immersive, l'ère informatique renverse le premier aspect tout en conservant le second. Dans les environnements multimédias, l'image prend le dessus, car son idéal est la transparence absolue et l'immersion dans la réalité virtuelle qui équivaut à un « déni de l'*ekphrasis* [37] ». Mais l'élimination totale du texte, écrit Bolter, ne tient pas compte du fait que même « les systèmes de réalité virtuelle reposent sur des couches successives d'écriture, de signes arbitraires sous forme de programmes informatiques [38]. »

La condition numérique, comme on pourrait paraphraser le travail de Jerome McGann, *est* donc la condition textuelle [39]. Tout est texte et toute image n'est jamais qu'une image pour nous. Même dans le modèle séquentiel, la distinction entre image et texte est dissoute en faisant du texte le seul mode d'existence des objets numériques. Il est donc toujours logique de parler ici d'*ekphrasis* opératoire mais il y a maintenant trois textes en jeu : le code, les textes des fichiers et le texte en tant qu'image tel qu'il apparaît à un lecteur humain. L'aspect performatif reste le même : le texte fait quelque chose qui est en fin de compte une image désormais augmentée par l'effet d'une sémiose secondaire qui a lieu non pas dans la machine, mais chez les humains.

La deuxième objection concerne la relation entre les concepts de « texte » et de « langue ». Elle semble avoir une portée extrêmement limitée : j'ai utilisé le terme « texte » pour parler des éléments de la pièce de Mon, du code de Meerhoff et, enfin, des données du fichier image. Ce sont tous des types de texte très différents, mais aucun d'entre eux n'est un langage qui, au sens plein du terme, a non seulement une syntaxe, mais aussi une sémantique et une pragmatique. Pourtant, le débat sur la question de savoir si le code peut prétendre être un langage au sens propre est complexe. Pour certains, comme Loss Pequeño Glazier, il n'y a pratiquement aucune différence entre les deux [40]. Le code, selon ce point de vue, peut donc également être un support poétique, un moyen d'expression. Pour d'autres, en revanche, toute signification qu'un tel code porte *pour nous* est simplement « parasite » des significations que nous lui associons, comme l'a affirmé Stevan Harnad, qui a récemment été abondamment cité dans la discussion sur la capacité des systèmes d'intelligence artificielle à produire du sens [41].

Pour ce dernier groupe, le langage artificiel du script n'est donc pas vraiment un langage à proprement parler. Florian Cramer fait écho à Harnad

36. C'est ainsi que Friedrich A. Kittler, peu après avoir déclaré qu'il n'y a pas de logiciel mais seulement du matériel, a pu exclure les images de synthèse de la classe des médias optiques en les déclarant essentiellement alphabétiques. Une image de pixels, écrivait-il, « trompe l'œil, qui est censé ne pas pouvoir différencier les pixels individuels, avec l'illusion ou l'image d'une image, alors qu'en vérité la masse des pixels, en raison de son adressage complet, s'avère être structurée plutôt comme un texte composé entièrement de lettres individuelles », in « Computer Graphics : A Semi-Technical Introduction », trad. Sara Ogger, *Salle grise*, 2, nº 2, 2001, p. 32.
37. Jay David Bolter, « *Ekphrasis*, Virtual Reality, and the Future of Writing », in Geoffrey Nunberg dir., *The Future of the Book*, Berkeley, University of California Press, 1996, p. 269.
38. *Ibidem*, p. 270.
39. Je me réfère au titre de Jerome J. McGann, *The Textual Condition*, Princeton University Press, 1991, mais l'idée que tout ce qui est numérique est compris comme un texte peut être trouvée dans Jerome J. McGann, *Radiant Textuality : Literature after the World Wide Web*, New York, Palgrave, 2001, p. 11.
40. Loss Pequeño Glazier, « Code as Language », *Leonardo* 14, nº 5, 2006, http://leoalmanac.org/journal/vol_14/lea_v14_n05-06/lpglazier.asp ; consulté le 7 avril 2024. Pour une version philosophiquement plus sophistiquée de cet argument, voir Juan Luis Gastaldi, « Why Can Computers Understand Natural Language ? The Structuralist Image of Language Behind Word Embeddings », *Philosophy & Technology* 34, nº 1, 2021, pp. 149-214.
41. Stevan Harnad, « The Symbol Grounding Problem », *Physica D : Nonlinear Phenomena* 42, nº 1-3, 1990, pp. 335-46. Pour les points de vue actuels, en particulier sur les grands modèles linguistiques, voir Emily M. Bender et Alexander Koller, « Climbing towards NLU : On Meaning, Form, and Understanding in the Age of Data », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, 2020, 5185-98 ainsi que Murray Shanahan, « Talking About Large Language Models » (arXiv, 11 décembre 2022), http://arxiv.org/abs/2212.03551 ; consulté le 7 avril 2024.

lorsqu'il qualifie les codes de programmation de « langages syntaxiques par opposition aux langages sémantiques ». Comme leur nom l'indique, les langages syntaxiques sont totalement dépourvus de sens, contrairement aux langages naturels, c'est-à-dire sémantiques. Cramer explique :

> Les symboles des langages de contrôle informatique ont inévitablement des *connotations* sémantiques, tout simplement parce qu'il n'existe pas de symboles auxquels les humains n'associeraient pas une certaine signification. Mais les symboles ne peuvent dénoter aucun énoncé sémantique, c'est-à-dire qu'ils n'expriment pas de sens en leurs propres termes [42].

Dans la mesure où la pragmatique est liée aux effets de sens, cela signifie également que le code n'est performatif que dans un sens technique, en tant que série de commandes exécutées selon des règles prédéfinies. Aucune de ces commandes n'est porteuse de sens en soi, qu'il s'agisse d'une référence au monde extérieur ou d'un système de signes dans le contexte de la communication. Le code est une syntaxe sans sémantique et il n'a de pragmatique que dans le sens abstrait de sa structure de commande [43].

Je suis prêt à admettre tout cela. En fait, c'est là mon point de vue pour la suite. En effet, dans la différenciation interne du numérique, cette notion limitée de texte ainsi que la relation du langage à l'image commencent à changer une fois que nous passons du paradigme séquentiel au paradigme connexionniste. En effet, dans les réseaux neuronaux, il n'y a pas de « premier texte » comme dans *They Lay* de Meerhoff ni de code écrit sous la forme d'une série de règles que nous pourrions inspecter et qui exécuteraient des commandes. Au lieu de cela, les données de départ passent par le réseau de connexions. Elles sont soit augmentées, soit diminuées à chaque étape, en fonction des poids formés. Enfin, les résultats sont transmis à la couche finale de neurones et additionnés pour produire une sortie unique. C'est le processus de base par lequel les réseaux neuronaux génèrent des prédictions à partir des données d'entrée. La sortie est donc le résultat d'un processus cumulatif, statistique et parallèle qui se déroule entre les nombreuses connexions du réseau, mais qui ne peut en aucun cas être considéré comme une *commande*.

Cependant, cela conduit à la conclusion curieuse selon laquelle, comparé au paradigme séquentiel, celui de l'algorithme classique qui est dépourvu de sémantique, le paradigme connexionniste n'a pas de structure de commande discernable et donc pas de pragmatique. Paradoxalement, cependant, la sémantique revient dans l'IA multimodale avec l'exemple de DALL-E 2. Et elle le fait en détruisant *la* distinction image/texte à un niveau plus profond que ne l'a fait la réduction des données d'image en texte dans le modèle séquentiel.

Je consacrerai la dernière partie de cet essai à suivre ce chiasme au cœur de la distinction séquentielle/connexionniste. Une première indication que le langage orienté vers le sens joue un rôle ici a été donnée par le texte d'entrée :

42. Florian Cramer, « Language », dans dir. Matthew Fuller, *Software Studies : A Lexicon*, Cambridge (Mass.), MIT Press, 2008, pp. 168-1699.
43. Cependant, il existe un débat animé sur l'utilité de parler d'une pragmatique des langages de programmation dans un sens plus large. J. H. Connolly et D. J. Cooke suggèrent que « les effets pragmatiques du programme en cours d'exécution… provoquent des changements dans l'état interne de l'ordinateur », in « The Pragmatics of Programming Languages », *Semiotica*, 151, 2004, p. 154. Benjamin Bratton suggère également que « le code est une sorte de langage exécutable. […] Dans ce sens, la "fonction" linguistique ne se réfère pas seulement à la compétence en matière de manipulation de symboles, mais aussi aux fonctions et aux effets du code exécuté dans le monde réel » : voir Benjamin Bratton et Blaise Agüera y Arcas, « The Model Is The Message », *Noema*, https://www.noemamag.com/the-model-is-the-message ; consulté le 12 juillet 2022.

après tout, l'intérêt de Dall-E est qu'il peut transformer un prompt en langage naturel autrement dit une description linguistique significative en un fichier image. Il s'agit là aussi d'une « peinture avec des mots », non pas en tant que représentation mais en tant que performance. DALL-E 2 peut donc raisonnablement être considéré comme un type d'*ekphrasis* opérationnelle : il agit comme un texte qui produit une image de manière informatique. Mais cette coordination du texte et de l'image ne peut se produire qu'en défaisant la distinction entre les deux et non par le biais du code, mais par quelque chose que l'on peut appeler la « sémantique artificielle ». Pour comprendre cela, nous devons à nouveau penser avec l'IA.

SÉMANTIQUE ARTIFICIELLE

L'IA multimodale est le nom donné à une nouvelle classe de réseaux neuronaux, dont DALL-E 2 est l'un des exemples les plus marquants. Ces modèles se distinguent par leur capacité à intégrer plusieurs types de données, telles que des images, du texte, de la parole, des données tactiles ou de localisation, et bien d'autres encore, afin d'accroître leurs capacités [44]. Alors que DALL-E 2 se concentre principalement sur la connexion d'images et de textes, il existe de nombreuses autres IA multimodales conçues pour traiter différents types de données, notamment le son ou les données de mouvement en 3D utilisées dans la conduite autonome. Il convient de noter que même des modèles apparemment monomodaux comme le GPT-4, qui génère principalement du texte, sont désormais formés à de multiples modalités [45]. Dans tous les cas, la caractéristique distinctive de ces réseaux est leur capacité à corréler et à traiter différents types de données. Par conséquent, ils dépassent les limites des anciens types de réseaux neuronaux qui étaient plus spécialisés et spécifiques à un support.

Dans le domaine des réseaux neuronaux, différentes « architectures » ont traditionnellement été conçues pour des tâches spécifiques. Certaines excellent dans le traitement de séquences temporelles, tandis que d'autres font preuve de performances supérieures dans le traitement d'informations spatiales. Cette division fait écho à l'argument de Lessing en faveur de la séparation des arts, et, en effet, certaines IA se révèlent mieux adaptées au traitement de textes, d'autres à celui d'images. Auparavant, deux architectures fondamentales, le réseau neuronal récurrent (RNN) et le réseau neuronal convolutif (CNN), représentaient les modèles de base dans ces domaines respectifs. Les CNN excellaient dans la génération d'images en raison de leur capacité à traiter efficacement des matrices bidimensionnelles, tandis que les RNN étaient plus adaptés à l'analyse textuelle, retenant des informations à partir de données ordonnées linéairement [46]. Ces réseaux étaient donc limités par leur association à un support particulier et intrinsèquement monomodaux.

44. Voir Paul Pu Liang, Amir Zadeh et Louis-Philippe Morency, « Foundations and Trends in Multimodal Machine Learning : Principles, Challenges, and Open Questions » (arXiv, 20 février 2023), https://arxiv.org/abs/2209.03430 et Cem Akkus et al. « Multimodal Deep Learning » (arXiv, 12 janvier 2023), http://arxiv.org/abs/2301.04856, consulté le 7 avril 2024.
45. OpenAI, « GPT-4 Technical Report », 2023, https://doi.org/10.48550/ARXIV.2303.08774 ; consulté le 7 avril 2024.
46. Voir Bajohr, « Algorithmic Empathy : Toward a Critique of Aesthetic AI », *Configurations*, Johns Hopkins University Press and the Society for Literature, Science, and the Arts, 2022.

C'est du moins ce qui s'est passé jusqu'en janvier 2021, date à laquelle OpenAI a dévoilé la première version, plus compacte, de Dall-E. Ce modèle peut transformer des informations textuelles en informations visuelles. Cependant, plutôt que de simplement assembler un RNN et un CNN, il a adopté une nouvelle approche, une architecture unique qui traite à la fois le texte et l'image, une véritable IA multimodale. Si Dall-E et son successeur DALL-E 2 sont toujours constitués de plusieurs réseaux neuronaux individuels qui fonctionnent en tandem, ils utilisent tous la même architecture, appelée Transformer, qui excelle dans le traitement des représentations condensées d'images *et* de textes [47].

Il convient de décortiquer la fonctionnalité de DALL-E 2, qui opère dans une phase d'entraînement et une phase générative (ou d'inférence). Dans la phase d'apprentissage, un modèle Transformer appelé CLIP (Contrastive Language-Image Pre-training) se voit présenter des centaines de millions d'images et les légendes correspondantes tirées d'Internet comme par exemple, une photo de chat avec la légende « c'est une photo de chat ». Il est ensuite entraîné à apprendre un « espace de représentation » dans lequel les images et les textes apparentés sont plus proches les uns des autres. Cette corrélation des informations sur les images et les textes est cruciale pour l'apprentissage de DALL-E 2, qui apprend à partir de l'espace de représentation établi par CLIP et s'en inspire pour créer son propre modèle interne, appelé « antériorité ». Cet « a priori » capture les propriétés statistiques des caractéristiques de haut niveau des données et forme une sorte d'échafaudage ou de fonction directrice que le processus génératif utilise pour produire des résultats. Le point central ici est que l'image et l'information textuelle ne sont pas stockées séparément : une fois corrélées par CLIP, elles font partie du même espace de représentation partagé utilisé par DALL-E 2.

La deuxième étape du fonctionnement de DALL-E 2 est la phase générative, au cours de laquelle un modèle distinct appelé GLIDE est activé. GLIDE exploite les données de corrélation stockées entre le texte et les images dans le modèle CLIP pour exécuter une opération inverse : au lieu de faire correspondre une image avec le texte correspondant, il synthétise une image qui s'aligne le mieux avec le prompt textuel fourni et ce par le biais d'un processus appelé « diffusion [48] ». Ce qui est important ici, c'est que GLIDE utilise l'espace de représentation de CLIP pour manifester les prompts de texte dans leurs contreparties d'images les plus probables. Ainsi, lorsqu'on lui présente un prompt tel que « un astronaute chevauchant un cheval dans un style photoréaliste », DALL-E 2, grâce à cette interaction de modèles collaboratifs, est capable de produire l'image d'un astronaute chevauchant un cheval rendue dans des détails photoréalistes. Cette capacité repose sur les apprentissages initiaux du modèle CLIP concernant les caractéristiques visuelles des « astronautes », des « chevaux » et du « style photoréaliste », ainsi que sur le pouvoir génératif de

47. Ashish Vaswani et al., « Attention Is All You Need », in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008. Voir aussi pour une explication étape par étape : Jay Allamar, « The Illustrated Transformer », 2018, https://jalammar.github.io/illustrated-transformer, consulté le 7 avril 2024.

48. Voir pour les modèles de diffusion Prafulla Dhariwal et Alex Nichol, « Diffusion Models Beat GANs on Image Synthesis » (arXiv, 1er juin 2021), https://doi.org/10.48550/arXiv.2105.05233 ; consulté le 7 avril 2024.

GLIDE qui synthétise ces concepts en une nouvelle composition visuelle. C'est ainsi que le poème de Dave Orr est né du prompt « un poème sur la singularité écrit dans une police à empattement ». Comme Dall-E 2 est stochastique et qu'il est destiné à produire des images plutôt que des textes, le résultat est flou et asémique mais il a clairement la *Gestalt* d'un poème. Ce qui est central dans toute cette opération, c'est que le modèle, comme le dit un interprète, « apprend le *lien sémantique* entre les descriptions textuelles des objets et leurs manifestations visuelles correspondantes [49] ». CLIP stocke les informations linguistiques et picturales dans le même espace de représentation. Le sens est le sens, quel que soit son support.

Cela est confirmé par le fait que les modèles multimodaux semblent parfois former des « neurones » uniques pour les concepts, indépendamment du fait que l'entrée soit visuelle ou verbale, parallèlement à ce qui a été supposé être des « cellules grand-mère » dans les neurosciences depuis au moins les années soixante [50]. Ce concept est né en réponse à la question de savoir comment les connaissances sont stockées dans le cerveau. Lorsque je vois une photo de ma grand-mère, cette reconnaissance est-elle le résultat d'une interaction complexe entre les régions du cerveau ? Ou y a-t-il *un* neurone spécifique qui s'allume, une cellule grand-mère ? En 2005, une étude neuroscientifique a suggéré que de tels neurones pourraient bien exister. Lorsque des images de l'actrice populaire Halle Berry étaient montrées à des sujets, une activité neuronale très localisée a été observée dans le lobe temporal médian. En outre, cette activité se produisait non seulement lorsque les sujets voyaient une photo de Berry, mais aussi lorsqu'ils voyaient un dessin d'elle et même la chaîne de lettres épelant « Halle Berry ». Cela a conduit les auteurs à suggérer que le cerveau pourrait utiliser un « code invariant, clairsemé et explicite » qui traite « une représentation abstraite de l'identité de la personne ou de l'objet montré » [51]. En d'autres termes, le cerveau pourrait encoder des concepts directement, de manière multimodale.

Un phénomène similaire a été constaté dans les « neurones » de CLIP, le modèle de DALL-E 2 qui coordonne le texte et l'image. En 2021, les chercheurs de l'OpenAI ont publié un article suggérant que les dernières couches d'un réseau CLIP entièrement entraîné présentent également quelque chose qui ressemble à une cellule grand-mère réagissant à des visages individuels. Il existe un neurone (l'article utilise Spiderman plutôt que Halle Berry) qui réagit également aux photos, aux dessins et aux textes qui se réfèrent à la même entité. Une photo de Spiderman et une chaîne de texte contenant son nom activeront le même neurone, tout comme une photo d'araignée, ce qui indique que ces neurones conceptuels sont regroupés sur le plan sémantique [52].

Pour être clair : la notion de neurones de grand-mère est très contestée. En neurosciences, cette interprétation est controversée et, en général, l'affirmation

49. Ryan O'Connor, « How DALL-E 2 Actually Works », *Assembly AI*, 19 avril 2022, https://www.assemblyai.com/blog/how-dall-e-2-actually-works/; consulté le 7 avril 2024.
50. Charles G. Gross, « Genealogy of the "Grandmother Cell" », *The Neuroscientist*, 8, nº 5, octobre 2002, pp. 512-518, https://doi.org/10.1177/107385802237175 ; consulté le 7 avril 2024.
51. R. Quian Quiroga et al., « Invariant Visual Representation by Single Neurons in the Human Brain », *Nature*, 435, nº 7045 (juin 2005) : 1102, 1106, https://doi.org/10.1038/nature03687 ; consulté le 7 avril 2024.
52. Gabriel Goh et al., « Multimodal Neurons in Artificial Neural Networks », *Distill*, 6, nº 3, 4 mars 2021, 10.23915/distill.00030, https://doi.org/10.23915/distill.00030 ; consulté le 7 avril 2024.

d'une certaine forme d'homologie entre le tissu cérébral réel et les réseaux neuronaux est au mieux une simplification excessive. En réalité, les choses sont plus confuses, comme le soulignent volontiers les auteurs de l'article CLIP. Malgré ces réserves, la notion de neurones grand-mère et celle d'espace de représentation partagé entre le texte et l'image semblent utiles pour mettre en évidence une tendance générale de l'IA multimodale. En ce qui concerne ses conséquences théoriques, et en particulier les conséquences pour la relation entre le texte et l'image, nous pouvons, dans l'esprit de la pensée avec l'IA, tirer déjà quelques conclusions même si les données empiriques sont incomplètes et doivent encore être discutées.

Si DALL-E 2, dont CLIP fait partie, encode ainsi texte et image dans les mêmes neurones ou dans le même espace de représentation, deux choses semblent s'ensuivre.

Premièrement, contrairement au modèle séquentiel, dans lequel le code était un système purement syntaxique avec une pragmatique limitée et aucune valeur sémantique, dans l'IA multimodale, la sémantique entre à nouveau en jeu. Je ne veux pas dire qu'il s'agit de sémantique au sens *plein*, qu'il s'agisse de l'« intention communicative » de la communication humaine explorée par la linguistique [53], ou de l'« être-en-situation » que la critique de l'IA d'Hubert Dreyfus, inspirée par Heidegger, présente comme la condition limitative pour des agents véritablement intelligents [54]. Mais il semble évident qu'en corrélant le texte et l'image au sein d'un système informatique unique dans l'IA multimodale, la différence entre le paradigme séquentiel et le paradigme connexionniste de la numéricité apparaît le plus clairement. On peut, en effet, avancer que les réseaux neuronaux et les modèles multimodaux en particulier, peuvent s'intéresser à quelque chose qui n'est peut-être pas du sens au sens plein de la communication humaine, mais que l'on ne peut pas non plus qualifier avec certitude de non-sens. C'est ce que j'appelle la sémantique artificielle et c'est ce qui rend les modèles d'IA si intéressants : ils ne portent pas seulement les connotations externes que nous projetons sur eux, comme l'a suggéré Cramer, mais ils génèrent également un certain type de sens inhérent grâce à la corrélation complexe du texte et de l'image au sein d'un système unique.

Il en découle un deuxième point. L'IA multimodale a pour effet d'abolir la distinction entre le texte et l'image. Les deux ne sont pas seulement corrélés dans le processus d'entraînement (même dépassés au niveau du système) ni liés à des représentations de texte ou d'image, mais identifiés [55]. D'un point de vue conceptuel, l'IA multimodale suggère une nouvelle position dans la tradition et l'ontologie de l'*ekphrasis* que j'ai décrite plus haut. L'interaction texte/image n'est plus à la base de toutes les théories traditionnelles, qu'elles soient représentatives ou performatives. La formulation de l'*ekphrasis* par l'IA multimodale suggère une identité structurelle entre le texte et l'image, les libérant de leur

53. Emily M. Bender et Alexander Koller, « Climbing towards NLU : On Meaning, Form, and Understanding in the Age of Data », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, op. cit.*, 5185-98, https://doi.org/10.18653/v1/2020.acl-main.463 ; consulté le 7 avril 2024.
54. Hubert L. Dreyfus, *What Computers Still Can't Do : A Critique of Artificial Reason*, Cambridge, MA, MIT Press, 1992.
55. En tant que telle, l'IA multimodale est plus qu'une remédiation, comme l'a suggéré Jay David Bolter, puisque ce terme conserve intacte la séparation des médias, Jay David Bolter, « AI Generative Art as Algorithmic Remediation », *IMAGE*, 37, nº 1, 2023, pp. 195-207.

fonction sémantique primaire. Il existe désormais, comme on pourrait l'appeler en citant Liliane Louvel un « tiers pictural multimodal [56] », le sens partagé dans le neurone artificiel, qui agit comme lieu de sémantique au-delà du mot et de l'image. Cela va à l'encontre de la crainte ekphrastique de la tradition formaliste de Lessing à Clement Greenberg qui prônait la séparation des médiums, mais cela explose aussi l'espoir ekphrastique de la lignée commençant avec Horace, fondée sur la transformation productive du genre. Ici, penser avec l'IA a donné naissance à une position véritablement nouvelle, et de grands modèles visuels tels que DALL-E 2 en constituent la mise en œuvre technique.

Enfin, un troisième point, comme je l'ai indiqué, le statut du langage change entre le paradigme séquentiel et le paradigme connexionniste. Les œuvres de Jasmin Meerhoff et de David Orr représentent chacune l'un de ces paradigmes, et constituent chacune un type d'*ekphrasis* opératoire – un texte qui produit une image. Mais alors que dans le cas séquentiel il y a une « pragmatique » sans sémantique, dans le cas connexionniste nous avons une « sémantique » sans pragmatique. Dans le premier cas, c'est le code qui « agit » sans porter le sens au-delà de sa simple valence symbolique dans un système d'opérations. Dans l'autre, c'est le modèle de poids qui « signifie » sans réaliser quoi que ce soit qui ressemble à un acte de langage. Le performatif se situe ici au début de la chaîne opérationnelle, dans la formulation de l'invite. Ainsi, le poème d'Orr signifie réellement ce qu'il montre à un niveau technique, totalement non-intentionnel, différemment de Meerhoff : il encode la description de lui-même *en lui-même*, soulignant une fois de plus que les images d'IA sont en effet quelque chose d'entièrement différent des œuvres classiques générées par le code.

CONCLUSION

J'ai rassemblé ici quelques idées sur la relation entre le texte et l'image dans le numérique et j'ai suggéré qu'avec l'avènement de l'apprentissage automatique stochastique sous la forme de réseaux neuronaux artificiels, il est nécessaire de diviser le domaine numérique en un sous-domaine séquentiel et un sous-domaine connexionniste. En outre, j'ai soutenu que seul le domaine numérique peut trouver ce que l'on pourrait appeler l'*ekphrasis* opérationnelle : les textes n'y représentent pas des images, mais les exécutent en les mettant en œuvre de manière informatique. Et correspondant aux approches connexionniste et séquentielle, il semble y avoir deux types distincts d'ekphrasis opératives, impliquant deux notions distinctes du langage : l'une mettant l'accent sur une dimension pragmatique, l'autre sur une dimension sémantique. Toutes deux, pour le répéter, sont très en deçà de la pleine signification de ces mots, mais ont néanmoins un lien raisonnable avec eux. Cependant, à l'encontre de

56. Liliane Louvel, *Le Tiers pictural*, Rennes, Presses Universitaires de Rennes, 2010.

l'orthodoxie selon laquelle les ordinateurs n'ont qu'une syntaxe sans sémantique, il existe au moins la possibilité que l'IA multimodale, dans ses neurones conceptuels, encode en fait du sens, un type de sémantique artificielle qui ne signifie pas tout à fait ce que les humains signifient, mais qui signifie quand même.

L'argument que j'ai avancé a donc une dimension à la fois concrète et méthodologique. D'une part, il sert une analyse esthétique de l'IA qui prend en compte le substrat technique de ses supports. Il s'agit d'un plaidoyer pour la multimodalité dans la discussion de ces œuvres. Elle montre qu'« il n'y a pas de médiums visuels », comme le disait W. J. T Mitchell, pour qui la séparation des médiums ignore toujours l'enchevêtrement des sens et la base linguistique de leur transmission [57]. Au même moment, nous n'avons ni Lessing ni Horace à suivre, mais quelque chose d'autre qui va au-delà de ces options. D'un autre côté, ce développement était également un exemple de la manière dont les *Critical AI Studies* pourraient non seulement penser sur ou contre, mais aussi *avec* l'IA. Le terme que j'ai proposé, l'*ekphrasis* opératoire, était dans ce cas moins destiné à ajouter une nouvelle dimension à un concept ancien et vénérable. Dans ce cas, le problème à étudier était le lien entre le texte et l'image, et l'interaction entre une métaphore technique et son utilisation humaniste.

Nous vivons une époque intéressante : sur le plan technique, les progrès de l'IA sont fulgurants et il y a un peu plus de trois ans, les phrases grammaticalement correctes générées par ordinateur étaient remarquables en elles-mêmes. Aujourd'hui, de simples descriptions génèrent des images. S'il ne faut pas se laisser prendre au piège de l'engouement pour l'IA et attribuer aux machines des caractéristiques telles que la conscience ou à ses constructeurs le statut de visionnaires pour qui les règles du fair-play ne tiennent plus, on ne peut pas non plus ignorer ces évolutions. Les catégories culturelles, philosophiques et esthétiques sont lentes à rattraper la réalité devant nos yeux et si la recherche peut les observer de loin ou s'impliquer directement, elle doit être ouverte à l'ajustement de ses catégories. L'*ekphrasis* opérative est l'un de ces ajustements.

Traduction Carla Marand

57. W. J. T. Mitchell, « There Are No Visual Media », *Journal of Visual Culture* 4, n° 2 (2005) : 257-66, https://doi.org/10.1177/1470412905054673 ; consulté le 7 avril 2024.