Title

Operative ekphrasis: the collapse of the text/image distinction in multimodal AI

Journal

Word & Image, 40(2)

ISSN

0266-6286

Author

Bajohr, Hannes

Publication Date

2024-04-02

DOI

10.1080/02666286.2024.2330335

Routledge
Taylor & Francis Group

Word & Image
A Journal of Verbal/Visual Enquiry

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/twim20

Operative ekphrasis: the collapse of the text/image distinction in multimodal AI

Hannes Bajohr

To cite this article: Hannes Bajohr (2024) Operative ekphrasis: the collapse of the text/image distinction in multimodal AI, Word & Image, 40:2, 77-90, DOI: 10.1080/02666286.2024.2330335

To link to this article: https://doi.org/10.1080/02666286.2024.2330335

Published online: 27 Jun 2024.

Operative ekphrasis: the collapse of the text/image distinction in multimodal AI

HANNES BAJOHR

Abstract This article discusses the implications of multimodal artificial intelligence (AI), including image generators such as DALL·E, for the traditional concept of ekphrasis. Using ekphrasis as an example of 'thinking with AI', it takes up the suggestion that in the digital realm ekphrastic relationships should be understood as performative rather than representational. Since with the introduction of modern AI the digital realm needs to be divided into a sequential part (classic algorithms) and a connectionist part (artificial neural networks), the article shows how the latter part ultimately tends toward a collapse of the text/image distinction in the technical system. Artificial neural networks both encode images and text as the same type of information, and they do so differently from the sequential model. Only in the context of multimodal AI, unlike in analogue or sequential paradigms, ekphrasis goes beyond the separation of or transition between text and image, but rather transcends this difference.

Keywords ekphrasis; artificial intelligence; visual poetry; Critical AI Studies

The burgeoning field of Critical AI Studies brings the perspective of the humanities to the ever-accelerating development of what is, broadly and inaccurately, called 'artificial intelligence' (AI).[1] As Rita Raley and Jennifer Rhee point out, it wilfully takes up the contested moniker of AI—which is often more of a marketing term than a technical description, for which 'machine learning' (ML) would be more apt—and treats it metonymically for a whole socio-economic culture of technology, thus 'engaging AI as an assemblage of technological arrangements and sociotechnical practices, as concept, ideology, and *dispositif*'.[2] In its most influential variety, Critical AI Studies responds to the fact that AI in the shape of stochastic (probability-based) ML has become a core element of the global flow of capital and its extractive tendencies as well as a central technology of surveillance and racial and economic exclusion, which is why this field is concerned with the political, economic, and ethical ramifications of these technologies.[3] An equally important part of Critical AI Studies is devoted to dissecting the conceptual and philosophical assumptions that underlie the design and use of ML applications, which still more often than not treat their 'data' as objective and neutral representations of the world.[4] If, as Philip Agre put it already thirty years ago, 'AI is philosophy underneath,'[5] critical work is needed to make explicit what is most often only implicit in actually existing AI systems. Often, this means, to quote German philosopher Hans Blumenberg, 'to destroy what is supposedly "natural" and convict it of its "artificiality"'[6]—for AI is often not considered artificial *enough*. It is in this very crucible that the humanities, equipped with their critical, historical, and conceptual awareness, find their relevance magnified. As Fabian Offert and Thao Phan put it, 'current-generation machine learning models require current-generation modes of (humanist) critique'.[7]

But this relationship between AI and the humanities goes both ways: if AI already *is* philosophy not yet articulated, we can also turn Agre's adage around—as humanists, we would be remiss if we did not also test our *own* concepts against the new phenomena that computer science and engineering throw at us. Consequently, humanistic practices must evolve to grapple with the questions incited by ML technology, and not only think about, and often against, but sometimes also *with* AI. This does not mean dropping the critical stance but rather extending it to both sides of the equation, and including humanistic concepts as an object of inquiry and potential revision in light of the questions raised by Critical AI Studies. In this paper, I demonstrate one example of such 'thinking with AI' by shining a new light on an age-old question of humanist inquiry, and one animating this journal—the relationship between word and image.

In what follows, I will develop some intuitions about this relationship, and ask how it may be changing with the shift from classical algorithms to current state-of-the-art ML. In particular, I am interested in so-called 'multimodal AI', among which large visual models such as DALL·E or stable diffusion may be the best known. To think with AI here is to test this technology's theoretical ramifications for a more traditional concept pertaining to the interaction of word and image, namely ekphrasis, which I broaden here to include the technical substrate of this interaction in the digital under the title 'operative ekphrasis'. Using this concept, I show that multimodal AI does away with the separation of mediums that is at the core of ekphrasis, as this technology can process both text and image as *one* type of data. In so doing, I use AI as what Daniel Dennett calls an 'intuition pump'[8]—a tool that allows us to clarify conceptual implications otherwise unseen.

In the first part of this article, I use examples from visual poetry to discuss three text/image media: (1) analogue, (2) 'sequentially' digital (classic computing) and (3) 'connectionistically' digital (stochastic ML). I will argue that with the advent of ML, the division between digital and analogue media needs to be subdivided, as AI operates differently from older computational paradigms. In the second part, I discuss how the rhetorical figure of ekphrasis provides a framework for ordering this new subdivision by interpreting code as performative. Finally, I draw two conclusions: first, that the classical opposition between text and image, on which the concept of ekphrasis is based, dissolves in multimodal AI; and second, that semantics nevertheless returns to the digital, which hitherto has been seen only as a matter of syntax. Taken together, these claims question both our aesthetic lexicon and our understanding of digitality. As such, they underscore the cross-disciplinary significance of Critical AI Studies, and show that the humanities, with the necessary care and without falling for hype and exaggeration, can benefit from thinking with AI.

Text and image in the digital

As good a way as any to start discussing the relationship between text and image is to turn to visual poetry, which by its very nature brings visuality and textuality into dialogue. Figure 1 shows a work by German concrete poet Franz Mon. It is taken from his cycle 'non tot' published in 1964, and it consists of several typewritten lines shaped like a diamond, or perhaps a sail. The lines of the upper half repeat the word 'non', those in the bottom half the word 'tot'. The lines grow progressively more compressed toward the centre of the page, partly obscuring one another. Figure 2 shows a visual poem by the contemporary German digital author Jasmin Meerhoff, taken from her 2022 collection 'They Lay'. Here, scraps of typeset text are arranged in a repetitive, undulating pattern that might suggest flames rising from some unseen fuel. What these letters spell is difficult to decipher—they are certainly letters, but in their collage-like configuration, they are even more divorced from linguistic meaning than Mon's already enigmatic 'non/tot'. To the uninitiated viewer, in any case, the two pieces speak a shared poetic language that brings together letters in constellations in which the visual quality of the page rivals or surpasses the poems' semantic meaning. These poems are to be looked at as images as much (if not more than) they are meant to be read as lines of text. Viewed next to each other, it seems that not much has changed in the roughly sixty years separating these two works.

Compare this with the piece shown in figure 3. It is the work of Dave Orr and, like Meerhoff's, it was created in 2022. Unlike the first two poems, however, it appears to be of a quite different make. Its centred text alignment gives it the air of a more traditional, or even naïve, poetic paradigm that predates the visual poetry of the other two pieces. Yet a second look reveals that while the title is clearly legible, if

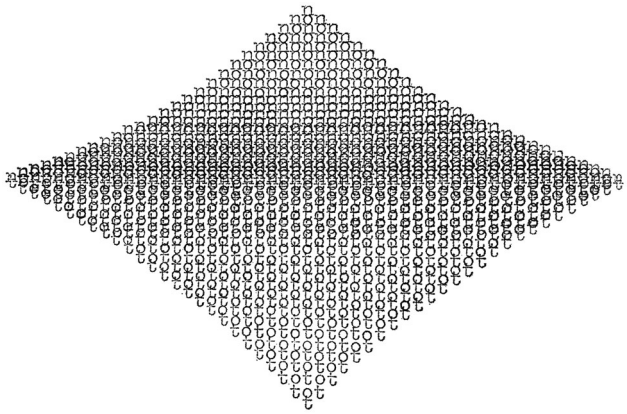

Figure 1. Franz Mon, 'non/tot III'; from *Lesebuch* (Neuwied: Luchterhand, 1964), 38. © 2024, ProLitteris, Zurich.

Figure 2. Jasmin Meerhoff, 'They Lay', 2022. Digital. nervousdata.com. Courtesy: Jasmin Meerhoff.

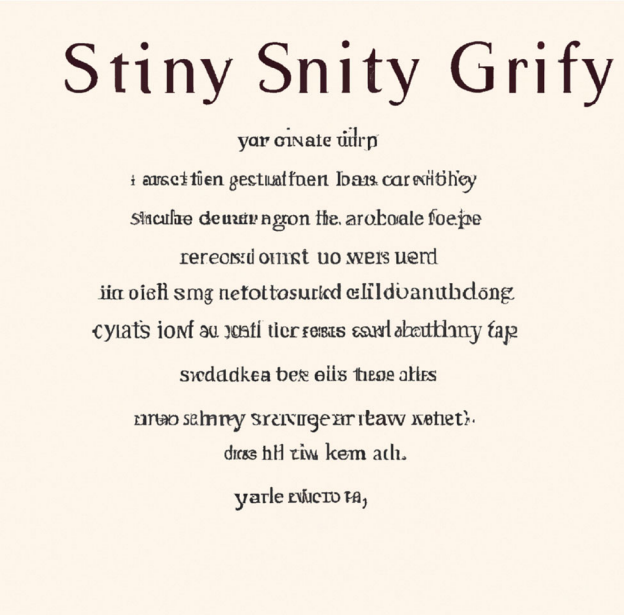

Figure 3. Dave Orr, 'Stiny Snity Grify', 2022. Digital. lesswrong.com. Courtesy: Dave Orr.

enigmatic—'Stiny Snity Grify'—the lines are in fact not simply nonsense; they are not even text. They have the character of what is often called 'asemic' writing, that is, writing that does not use words but merely the semblance of words. If, as Peter Schwenger puts it, the 'visual and muscular aspects of writing are generally obscured by the primacy of writing's communicative function', then an asemic text 'does not attempt to communicate any message other than its own nature as writing',[9] including its visual character. In this sense, Orr's poem, too, could be classified as 'visual', albeit from a divergent perspective compared with the other two—instead of making a poem by using text to create an image, it uses an image to create a poem that looks like text.

As instances of visual poetry, commonalities between the three works can be identified. What interests me here, however, is what sets them apart—and this is in no small part their technical substrate, their, as Katherine Hayles calls it, 'media-specificity'.[10] For all three use radically different technologies, and all these technologies imply radically different relationships between text and image.

Perhaps unsurprisingly, the two examples from 2022 use digital technology, while Mon's 1964 work was created by analogue means—with an Olympia Monica mechanical typewriter, to be exact, on which he produced much of his concrete and visual poetry. Figure 4 shows a section from the Monica's manual with a sample of its signature typeface. In contrast, Jasmin Meerhoff created her poem digitally, by writing a MacOS shell script (figure 5). When executed in the command line, the script tells the open-source application Imagemagick to do two things: first, to cut a single image file containing a line of text from a scanned page into small pieces (lines 12–20 in the script); and second, to collage those pieces into the shape that makes up the poem (lines 23–27). The wavy appearance is the result of using a sine function to arrange the pieces by specifying the amplitude and frequency of the waves (line 4). This is all done automatically, and Meerhoff's script is freely available online,[11] enabling anyone to make a potentially endless stream of visual poems.

Dave Orr's piece is also produced by digital means, but in a very different way. It was created using an 'artificial intelligence', or more precisely, a complex ML algorithm that is implemented as a neural network. The neural network in this case is called DALL·E, a product by the company OpenAI, best known for its text-generation model ChatGPT.[12] DALL·E, currently in its third version, is a large visual model with a text-to-image capability, and it is only one among a growing number of them, such as StabilityAI's Stable Diffusion, Google's Imagen, or Midjourney.[13] These systems take a natural language description (the 'prompt') as an input and generate an image as an output, producing a visual representation of the content of the text. In the case of DALL·E 2—which was used to produce Orr's poem—this is done via an interface that consists of a single text box for the input prompt (figure 6).[14] For Orr's poem, the prompt was 'a poem about the singularity written in a serif font'.[15]

It is worth noting that DALL·E typically does not generate texts. This 'poem' emerged when Orr was examining the model, and it appeared as part of a blog post about the system. As far as I am aware, it was never meant to be published as a literary work, and the fact that Orr is Google DeepMind's director of engineering and does not, to my knowledge, consider himself a poet supports this impression. Indeed, AI image generation is famously bad at producing text, and next to mangled hands, garbled writing is (or was until recently) the most prominent tell-tale sign that a picture is in fact AI generated.[16] What DALL·E is usually meant to produce are images—either photorealistic or stylized—all of which have in common that they are the result of an input text. Figure 7 shows a more typical example from the developer's website. The prompt 'An astronaut riding a horse in photorealistic style' results in an image of just that. As there is nothing but the textual prompt for users to steer the image generation, a veritable 'promptology' has established itself since the popularization of large visual models. By finessing the input text, adding more descriptions of style or atmosphere, it is possible to nudge the result in one direction or another. Apart from a *Prompt Book*,[17] there is now even a website on which particularly useful prompts are sold for small sums.[18]

The three discussed works each embody different poetic and technological paradigms, which can be categorized in

Figure 4. Excerpt from the Olympia Monica's manual (model SM7).

several different ways: first, as a type of text/image interaction in the broader genre of visual poetry; and second as the result of analogue (Mon) or digital technology (Meerhoff, Orr). However, it is possible to break down the digital technology into two subcategories: third, classical algorithms and modern AI, which I will discuss below under the rubric of sequential and connectionist paradigms respectively. For now, it is sufficient to note that the digital realm is not a monolith but instead a landscape of varied subdomains.

All three classifications connect text with images or image-like structures, but they do so in distinct ways. Visual poetry does this by its very nature: it creates images through the arrangement of text. However, only the two digital works do so at the level of technical substrate. Here, the text and the resulting text/image stand not simply in a mimetic relationship (text that, once arranged, looks like an image) but in a causal one (text that, as part of a transformative process, brings about an image). This is what I call *operative ekphrasis*. But while in the second case—the classic computer code—a purely syntactic code language is responsible for the process of making an image, only in the last case—the AI model—is there also a semantic element; it is this element that ultimately threatens to dissolve the distinction between text and image altogether. The remainder of this article will be spent unpacking these claims. They bear significant implications for how we interpret and understand these works and their differences, and it is an example of what an aesthetics of AI could be that takes its technical substrate seriously. To illustrate why, I need to elucidate to some degree how these technologies operate.

Sequential and connectionist paradigms of AI

For the subdivision of digital technology, I have proposed the terms 'sequential' and 'connectionist'.[19] The sequential paradigm denotes the dominant style of operating computers since Alan Turing's (conceptual) invention of the Universal Machine in 1936 and, after earlier models had been built, John von Neumann's (actual) implementation of the 'stored program' concept in the EDVAC (electronic discrete variable automatic computer) architecture in 1945 (built in 1949)[20] that by and large is still used today. It is characterized by the classical algorithm, laid down in a programming language of sequentially executed steps. Meerhoff's cut-up script belongs in this category, as do most of the programs on a typical computer. For instance, the command 'read -p' in line 4 (figure 5) requests a user input that will be stored in the variables 'am' and 'fm', which later designate the amplitude and the frequency of the poem's waves. Importantly, these lines are executed one after another and in a deterministic manner. Every time it is run, the program will go through the same, predictable commands. Because one can inspect the algorithm by reading the explicitly stated rules, this paradigm has, in principle, a high degree of transparency to human readers.

The sequential paradigm differs greatly from the newer digital mode of operation that I call connectionist, which is what is usually meant by AI today—deep learning, which is a subset of stochastic ML methodologies that uses multilayered artificial neural networks to model complex patterns in data. Loosely inspired by the way individual neurons in the brain repeatedly forge paths to perform higher level functions, current deep neural networks are made up of interconnected units, often referred to as 'neurons', which are linked by 'synapses'. (It is important to note, however, that this is a highly idealized affair and should not be confused with actual brain structure.) In these computational models, each neuron receives and processes incoming data, calculates a weighted sum based on its input, and then typically applies a non-linear

```
#!/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
```

Figure 5. The shell script for Jasmin Meerhoff’s ‘They Lay’.

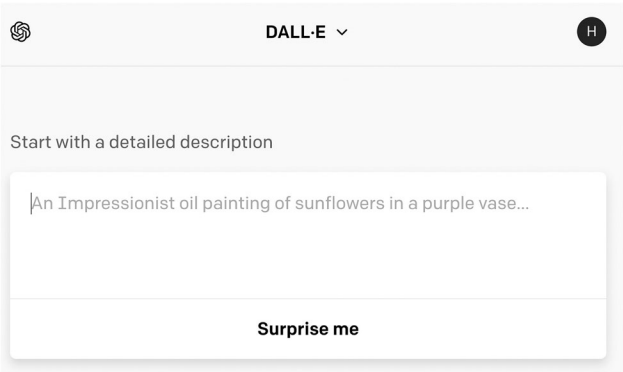

Figure 6. The interface of DALL·E with a text box for inputting the prompt.

activation function to determine its output in a process called ‘forward propagation’. The discrepancy between this output and the desired or known correct data is then measured using a loss function. Subsequently, an optimization algorithm—typically a variant of ‘gradient descent’—is used to adjust the weights and biases across the network to minimize this loss, a process known as ‘backpropagation’. The primary aim of training a deep neural network is to refine these parameters so that the model can generalize effectively, extrapolating from the training dataset to predict outcomes or classify new instances accurately. Put differently, the network identifies underlying patterns in the training set, fits a mathematical function to these data points, which then serves as a model for interpreting unseen data.[21]

Figure 7. A DALL·E-generated image for the prompt 'An astronaut riding a horse in photorealistic style'.

Consider a practical example involving image generation. Given a large enough dataset of human faces, a neural network can process this dataset to learn its inherent patterns, structures, and variations. These learned characteristics can then be applied to generate entirely new images of faces, which, despite being completely novel, will appear strikingly similar to real human faces. Because of the statistical nature of the AI, these faces are neither collages of face parts, nor mere linear composites of all the faces known to the model. Rather, and metaphorically speaking, the network learns *faceness*, the *Gestalt* of faces, and is able to recreate it in a way that does not repeat the individual inputs.[22] This is the principle of the well-known website thispersondoesnotexist.com, which presents a completely new and unique but artificial portrait of a face every time it is refreshed.

The AI model resulting from this training process implements complex nonlinear functions. What is central, now, is that a neural network cannot be translated back into a deterministic and exact higher level algorithm, as the model merely describes the connection strengths between the 'neurons' in the so-called weight model. While of course neural networks are still implemented in a von Neumann machine—and not, say, in an analogue or quantum computer—and are thus still digital, they nevertheless follow a radically different conceptual framework than the sequential model. For unlike the sequential paradigm, whose logic is laid out step by step, the connectionist paradigm follows a stochastic rather than a purely deterministic logic—in other words, the learned 'knowledge' is embedded in the network's structure and its weights, which represent the strength of connections between the artificial neurons. As a result, while it is technically possible to 'read' the values of the weights in a trained neural network, these numbers do not translate into a sequence of comprehensible instructions or steps in the same way that traditional code in a programming language does.[23]

Evidently, these are two very different models of computation that we nevertheless call 'digital'. Meerhoff's work, produced by a classical algorithm, was an example of the sequential, Orr's 'poem', produced by a neural net, of the connectionist paradigm. This technical introduction matters. For this nested distinction—that between analogue and digital, and that, within the digital, between the sequential and the connectionist paradigm—generates different relationships between text and image. What characterizes both digital forms, but not the analogue one, is what I want to call operative ekphrasis.

Representational and performative notions of ekphrasis

The concept of ekphrasis is one of the most-discussed terms in visual theory, literary criticism, and classics for describing the relationship between text and image. It has, as Ruth Webb noted, become a theoretical genre unto itself, evoking 'a network of interlocking questions and interests, from the positivist pursuit of lost monuments described in ancient and medieval ekphrasis to the poststructuralist fascination with a textual fragment which declares itself to be pure artifice, the representation of representation'.[24] But it has repeatedly been pointed out that in its original meaning, ekphrasis was a much broader category and signified a rhetorical device for generating vivid, sensory descriptions in oratory. A word from the rhetor's education, it was used to describe the act of clearly conjuring up something in the mind's eye of the audience—of transforming them, as Nikolaos of Myra put it in the second century CE, from listeners to spectators.[25]

This early meaning already implies a media anthropology in which the auditory and visual senses become functionally interchangeable. The ancient use did not particularly attend to the description of visual artworks, as Webb stresses.[26] It was only later, in the nineteenth and most emphatically in the twentieth century, that the term 'ekphrasis' became restricted to *literary* representations of a real or, in the case of what John Hollander has later called 'notional ekphrasis', an imaginary work of art.[27] Nevertheless, both interpretations persist in different guises to this day, so that, sampling the past five decades, definitions of ekphrasis have ranged from 'any description of anything visual'[28] to, more specifically, 'the poetic description of a pictorial or sculptural work of art'.[29]

Staying with the broader and, I think, more philosophically generative definition, it makes sense that James Heffernan has

emphasized the *representational* quality of ekphrasis by defining it as 'the verbal representation of visual representation'.[30] Tamar Yacobi has underscored this view by suggesting that ekphrasis is 'representation in the second degree' by specifying representation as repetition in a different mode: 'What was originally an autonomous image of the world becomes in ekphrastic transfer an image of an image, a part of a new whole, a visual *inset* within a verbal *frame*'.[31] We can see here already that this characterization, and the rhetoric of inset and frame, point to the tension at the core of ekphrasis: the concept articulates either an equivalence of or a competition between language and image—one imitating the other is either a successful enterprise or the recipe for disappointment. Thus, as W. J. T. Mitchell has noted, ekphrasis can be part of a hopeful or a fearful ontology of text/image interaction. It is either something with an almost utopian potential for transformation—from the visual to the verbal and back again, as the ancients had it—or a most blatant impossibility, which therefore needs to be prohibited aesthetically: making the visual absolutely verbal can never actually happen, and in fact it must not.[32] Lessing's *Laocoön*[33] argued for the incompatibility of language's temporal structure (ideal for depicting action) with painting's spatial makeup (best suited for depicting objects) and is, to Mitchell, the 'classic expression of ekphrastic fear'.[34]

This analysis of the immanent characteristics of the mediums involved in the metaphor of 'painting with words' (Horace) entails a critique of mimetic representation as the defining linchpin of ekphrasis. It has been taken up again in the last decade, and the focus on representation is replaced by a focus on *performance*: what is it that ekphrasis *does*, without saying that this doing must be imitative? Renate Brosch has thus suggested a new definition: 'ekphrasis is a literary response to a visual image […] emphasizing the performative instead of the mimetic'.[35] This performative interpretation of ekphrasis has several advantages, the main one being that by passing over its mimetic dimension, one can suspend the decision about its hopeful or fearful interpretation. Instead of understanding it as either an equivalence of or a competition between art forms—as a successful or unsuccessful relationship of representation—it simply places them in a consecutive and causal relation.

My reason for discussing visual poetry, which is not in the traditional, representational sense ekphrastic, is that it nevertheless can be understood as a performative ekphrasis: it is text 'doing' image. But beyond that, I would like to extend this notion by mobilizing the performative definition of ekphrasis for digital media in general. With a different emphasis, Brosch also brings ekphrasis to the digital, arguing that it becomes important in a digital media ecology that is inundated with images while also drowning in text—contradicting such doomsday predictions that saw the demise of reading. However, I will tweak her use of the word 'performative' to talk about ekphrasis in its media-specificity in the digital. Instead of literary 'responses', which are not themselves digital events, I want to understand the performance of ekphrasis as a *computational operation that correlates text and image*.

Operative ekphrasis

With the performative notion of ekphrasis in mind, let me return to the three visual poems, the division between analogue and digital, and the subdivision between sequential and connectionist. All three works embody specific ways of using language to create an image. In this sense, they are all ekphrastic in their cumulative effect: producing visual constellations through text. That alone, however, is not yet what I call operative ekphrasis. It is only really possible for text actively and causally to bring forth an image in the digital works, not in the analogue one.

In the analogue—in Mon's typewriter poem—the text may of course 'produce' an image. Yet this production is not performative on the operative level, but rather a perceptual aftereffect of a manual arrangement. In 'non/tot', it is the writer's bodily actions—his hand movements on the typewriter, his exerting force onto the keys—that lead to what we are compelled to describe as the 'image' of the text. This visual structure is the result of work, that is, a causal chain of mechanical forces that are not themselves textual. There is only one text here, the one on the page; it does not, strictly speaking, perform anything.

This is different in the digital works. In Meerhoff's piece, there are now two texts—the one on the page and the one that actually produces that text, the code. This is a textual performance in a *computational* sense: an operation the first text carries out to produce the second text effectively. It does so not as mechanical work, as in Mon, but as the manipulation of information, which is itself textual in nature. This is not a new insight, of course, and scholars like Espen Aarseth have built entire theories around this duality of text,[36] while Katherine Hayles has argued that 'electronic text is more processual than print, it is performative by its very nature'.[37] It is precisely this performativity of the interplay between the first text—the code—and the second text—the final constellation-image—that I call operative ekphrasis. It means understanding ekphrasis not as representation but as performance; not as imitating an image through text, but as text effectively bringing about an image. As such, it is truly 'words painting a picture'—but as an operation of manipulating symbolic information rather than figurative representation.[38]

Two remarks are necessary here that address possible objections to this notion of operative ekphrasis. First, it is easy to note that what is 'painted' here is not in fact a *text* image.

Meerhoff's work may use text (the code) to create an image composed of text (the work), but technically, the result is an image file, not a text file. Its content, once put up on a computer screen, only registers as text for humans, but not for machines. It is a bitmap image, a grid of pixels with different colour values, and as such it is human-readable but not machine-readable. Without a process of optical character recognition, the computer would not even register it as text.

The response to this objection is to note that the image, too, is, on a lower level, constituted textually: for image files are encoded alphanumerically. It is only through the translation of this text into the pixel matrix of a screen by means of a codec that it actually becomes an image.[39] This argument was levelled for describing digital ekphrasis as early as 1996, when media theorist Jay David Bolter declared that if the tradition of text/image interaction had been predicated both on the superiority of the word to the image as well as a metaphysics of presence that hoped to get to the thing itself through immersive description, the computer age reverses the first aspect while retaining the second. In multimedia environments, images take the lead, as their ideal is absolute transparency, and the immersion of virtual reality that amounts to a 'denial of ekphrasis'.[40] Yet a complete elimination of text, Bolter wrote, was oblivious to the fact that even 'virtual reality systems rest on layer after layer of writing, of arbitrary signs in the form of computer programs'.[41]

The digital condition, then, as one could paraphrase Jerome McGann's work, *is* the textual condition.[42] In the digital, everything is text, and every image is always only image-for-us. Even in the sequential model, the distinction between image and text is dissolved by making text virtually the only mode of existence for digital objects.[43] It thus still makes sense to speak of operative ekphrasis here, except that now there are three texts involved—the code, the (alphanumerically encoded) image file, and the text-as-image as it appears to a human reader. The performative aspect remains the same: text does something that is ultimately an image—which is now augmented by the effect of a secondary semiosis that takes place not in the machine but in humans.

The second objection regards the relationship between the concepts of 'text' and 'language'. It seems to have an extremely limited scope: I have used the term 'text' to speak of the elements in Mon's piece, of Meerhoff's code, and finally of the data in the image file. These are all very different kinds of text but none of them is language in the full meaning of the word, which not only has a syntax, but also a semantics and a pragmatics. Still, the debate as to whether code can claim to be a language in the proper sense is complex. For some, such as Loss Pequeño Glazier, there is practically no difference between the two.[44] Code, in this view, can thus equally be a poetic medium, a means of expression. For others, however, any meaning such code carries *for us* is simply 'parasitic' on the meanings we associate with it, as Stevan Harnad famously argued, and which has recently been reemphasized in the discussion about the AI systems' ability to produce meaning (a point to which I return in a later section).[45]

For the latter group, the artificial language of the script, then, is not really a proper language at all. Florian Cramer echoes Harnad when he calls programming codes 'syntactical languages as opposed to semantic languages'. As the name suggests, syntactical languages are utterly devoid of meaning, unlike natural, that is, semantic languages. Cramer explains:

> The symbols of computer control languages inevitably do have semantic *connotations* simply because there exist no symbols with which humans would not associate some meaning. But symbols can't denote any semantic statements, that is, they do not express meaning in their own terms.[46]

Insofar as pragmatics is tied to meaning-effects, this also means that code is performative only in a technical sense—as a series of commands that are executed according to predefined rules. None of these commands in themselves carries meaning, be it understood as reference to the outside world or a system of signs within the context of communication. Code is syntax without semantics; and it has a pragmatics only in the abstract sense of its command structure.[47]

I am willing to admit all this. In fact, this is my point moving on. For in the internal differentiation within the digital, this limited notion of text as well as the relationship of language to image begins to change once we turn from the sequential to the connectionist paradigm. In neural networks, there is no 'first text' as there was in Meerhoff's 'They Lay', no code that is written as a series of rule steps we could inspect and which, when executed, would perform commands. Instead, seed data is passed through the network of connections; it is either increased or decreased at each stage, depending on the trained weights. Finally, the results are output at the end layer of neurons, and summed together to produce a single output. This is the basic process by which neural networks generate predictions from input data. The output, then, is the result of a cumulative, statistical, and parallel process that takes place between the many connections of the network, but which cannot in any plausible way be thought of as command-like.

However, this leads to the curious conclusion that compared with the sequential paradigm—the classical algorithm, which is devoid of semantics—the connectionist paradigm has no discernible command-structure and therefore no pragmatics. Paradoxically, however, semantics returns in multimodal AI such as DALL·E. And it does so by collapsing the image/text distinction on a deeper level than did the reduction of image data to text in the sequential model.

I will spend the final part of this article following this chiasmus at the heart of the sequential/connectionist distinction. A first hint that meaning-oriented language plays a role here was given by the input text: after all, the whole point of DALL·E is that it can turn a natural language prompt—a meaningful linguistic description—into an image-file. This, too, is a 'painting with words', again, not as representation but as performance. DALL·E must thus also reasonably be called a type of operative ekphrasis: it acts as a text that computationally produces an image. But this coordination of text and image can only happen by undoing the distinction between them, and not through code but through something that may be called 'artificial semantics'. To understand this, we must again think with AI.

Artificial semantics

Multimodal AI is the name given to a new class of neural networks. The distinguishing feature of these models lies in their ability to integrate multiple data types, such as images, text, speech, tactile or location data, and more, to increase their performance.[48] A distinction can be made between multimodal AIs in which the input and output are of different modalities and those in which the inputs or outputs themselves are multimodal.[49] While DALL·E and other text-to-image models belong to the first type and primarily focus on converting one modality into another—text into image—multimodal AIs of the second type are designed to process different data types at once as enriched information type. GPT-4, which generates text, is now trained on multiple modalities to boost performance,[50] and with Gemini, Google introduced a large multimodal model that combines image, audio, video, and text data from the outset, as does OpenAI's newest system, GPT-4o.[51] In both cases, what distinguishes these networks from older models is their ability to correlate and process various types of data. Consequently, they transcend the limitations of older neural network types that were more specialized and medium-specific.

In the realm of neural networks, different 'architectures' have traditionally been tailored for specific tasks. Some excel at handling temporal sequences, while others demonstrate superior performance in processing spatial information. This division parallels Lessing's argument for the separation of the arts, and indeed certain AIs prove better suited to process text, others images. Previously, two fundamental architectures, the recurrent neural net (RNN) and the convolutional neural net (CNN), represented the core models in these respective domains. CNNs excelled in generating images due to their ability to handle two-dimensional matrices effectively, while RNNs were more suitable for textual analysis, retaining information from linearly ordered data.[52] Hence, these networks were constrained by their association with a particular medium and inherently *uni*modal.

For most users, this was the situation at least until January 2021 when OpenAI unveiled the inaugural, more compact, version of DALL·E. This model could transform textual into visual information. Rather than simply stitching an RNN and a CNN together, however, it adopted a new approach: a single architecture that handles both text and image, a truly multimodal AI. While DALL·E and its successors DALL·E 2 (2022) and DALL·E 3 (2023) still consist of several individual neural nets that work in tandem, they all utilize the same architecture, called the Transformer, which excels at dealing with condensed representations of images *and* text.[53]

It is worth unpacking the functionality of DALL·E, which operates in a training and a generative (or inference) phase. In the training phase, a transformer model called CLIP (contrastive language–image pre-training) is fed hundreds of millions of images and their associated captions taken from the internet—for example, a photo of a cat with the caption 'this is a photo of a cat'. Using a technique called contrastive learning, it is then trained to produce a single shared embedding space on which different modalities are mapped, so that related images and texts are closer within this space. The result is a very large model that correlates image information to text information.

This correlation of image and text information is crucial in the training of DALL·E. It learns from the embedding space established by CLIP, and builds upon it to create its own internal model called a 'prior'. This 'prior' captures the statistical properties of the high-level features in the data and forms a kind of scaffold that the generative process uses to produce outputs. The central point here is that image and text information are not stored separately; once correlated by CLIP, they become part of the same, shared representation space used by DALL·E, and are stored as *one* type of data.

The second step in DALL·E's operation is the generative phase, in which a separate model called GLIDE (guided language to image diffusion for generation and editing) is activated. GLIDE leverages the stored correlation data between text and images in the CLIP model to execute a reverse operation: rather than matching an image with corresponding text, it synthesizes an image that best aligns with the provided text prompt, and it does so through a process called 'diffusion'.[54] What is important here is that GLIDE uses CLIP's representation space to manifest text prompts into their most probable image counterparts. Thus, when presented with a prompt like 'an astronaut riding a horse in photorealistic style', DALL·E, through this collaborative model interplay, is able to output an image of an astronaut astride a horse, rendered in photorealistic detail. This ability hinges on the initial learning from the CLIP model about the visual characteristics of 'astronauts', 'horses', and 'photorealistic style', and the generative power of GLIDE to synthesize these concepts into a novel visual composition. It is

in this way that the prompt 'a poem about the singularity written in a serif font' resulted in Dave Orr's poem. Because DALL·E is stochastic, and because it is meant to output images rather than texts, the result is blurry and asemic, but it clearly has the *Gestalt* of a poem. What is central in this whole operation is that the model, as one interpreter puts it, 'learns the *semantic link* between text descriptions of objects and their corresponding visual manifestations'.[55] CLIP stores linguistic and pictorial information in the same representation space—meaning is meaning regardless of its medium.

To speak of 'meaning' here—be it understood as reference to the world or the communicative intent of speakers—may come as a surprise. After all, semantics was the absent dimension of the sequential paradigm, of classic code as a purely syntactical language not grounded in any connection to reality. Yet precisely because the connectionist paradigm in the shape of multimodal models correlates with different types of data, it might also be a contender for a limited, a 'dumb' kind of meaning.[56] This is borne out by the fact that multimodal models sometimes appear to form single 'neurons' for concepts independent of whether the input is visual or verbal, paralleling what have been hypothesized as 'grandmother cells' in neuroscience since at least since the nineteen sixties.[57] This concept arose in response to the question of how exactly knowledge is stored in the brain. When I see a picture of my grandmother, is this recognition the result of a complex interaction of brain regions? Or is there *one* specific neuron firing, a grandmother cell? In 2005, a neuroscience study suggested that such neurons may indeed exist. When subjects were shown images of popular actor Halle Berry, a highly localized neural activity was observed in the medial temporal lobe. Moreover, this activity occurred not only when subjects saw a photo of Berry but also when they saw a drawing of her and even the string of letters spelling out 'Halle Berry'. This led the authors to suggest that the brain may use an 'invariant, sparse and explicit code' that processes 'an abstract representation of the identity of the individual or object shown'.[58] In other words, the brain may encode concepts directly, in a multimodal fashion.

A similar phenomenon was found in the 'neurons' of CLIP, the model in DALL·E that coordinates text and image. In 2021, OpenAI researchers published a paper suggesting that the later layers of a fully trained CLIP network also show something like a grandmother cell responding to individual faces. There is a neuron—the paper uses Spiderman rather than Halle Berry—that also responds to photos, drawings, and text that refer to the same entity. A picture of Spiderman and a string of text with his name will both activate the same neuron, as does a picture of a spider, indicating that these conceptual neurons are clustered semantically.[59]

To be clear: the notion of grandmother neurons is very much contested—in neuroscience, this interpretation is controversial, and in general the claim of some kind of homology between actual brain tissue and neural networks is at best an oversimplification.[60] In reality, things are messier, as the authors of the CLIP paper readily point out. Despite these caveats, however, the notion of grandmother neurons—and that of the shared representation space of text and image—seems useful for highlighting a general tendency of multimodal AI. When it comes to its theoretical consequences, and in particular to the consequences for the relationship between text and image, we can, in the spirit of thinking with AI, already draw some conclusions even if the empirical data is incomplete and still in need of discussion.

If DALL·E, of which CLIP is a part, thus encodes text and image in the same neurons or in the same representation space, two things seem to follow.

First, unlike the sequential model, in which code was a purely syntactic system with a limited pragmatics and no semantic value, in multimodal AI, semantics comes back into play. I do not want to say that this is semantics in the *full* sense—be it the 'communicative intent' of human communication that linguistics explores,[61] or the 'being-in-a-situation' that the Heidegger-inspired AI critique of Hubert Dreyfus sets up as the limiting condition for truly intelligent agents.[62] But it seems clear that by correlating text and image within a single computational system in multimodal AI, the difference between the sequential and the connectionist paradigm of digitality shows itself most clearly. For one can make the argument that neural networks, and multimodal models in particular, may indeed be concerned with something that may not be meaning in the full sense of human communication, but cannot be confidently labelled non-meaning either. This 'dumb' meaning is what I call artificial semantics and it is what makes AI models such interesting artifacts: they not only carry the external connotations we project on them, as Cramer suggested, but also generate a certain type of inherent meaning through the intricate correlation of text and image within a single system.

From this follows a second point. The effect of multimodal AI is to collapse the distinction between text and image. Both are not only correlated in the training process but, on the system level, surpassed—not bound to either text or image representations, but a shared third.[63] Put conceptually, multimodal AI suggests a new position in the tradition and ontology of ekphrasis I described earlier. No longer the text/image interaction that underlies all its traditional theories, be they representative or performative, multimodal AI's formulation of ekphrasis suggests a structural identity between text and image on a higher level, relieving them of their primary semantic function. There is now, as one could call it with Liliane Louvel, a 'multimodal pictorial third'[64]—the shared meaning in the artificial neuron—that acts as locus of semantics beyond word and image. This flies in the face of the ekphrastic fear of the formalist tradition from Lessing to Clement

Greenberg that advocated for the separation of mediums, but it also explodes the ekphrastic hope of the lineage starting with Horace, based on the genre's productive transformation. Here, thinking with AI has yielded a genuinely new position, and large visual models such as DALL·E figure as its technical implementation.

Finally, a third point. As I have indicated, the status of language changes between the sequential and the connectionist paradigms. Jasmin Meerhoff's and David Orr's works each represent one of these paradigms, and each constitutes a type of operative ekphrasis—a text that produces an image. But whereas in the sequential case there is a 'pragmatics' without semantics, in the connectionist case we have a 'semantics' without pragmatics. In the first instance, it is the code that 'acts' without carrying meaning beyond its mere symbolic valence within a system of operations; in the other, it is the weight model that 'means' without carrying out anything that resembles a speech act. The performative here stands at the beginning of the operational chain, in formulating the prompt. Thus, Orr's poem really means what it shows—on a technical, fully non-intentional level—in a way that Meerhoff's does not: it encodes the description of itself *within* itself, highlighting once more that AI images are indeed something entirely different from classic code-generated works.

Conclusion

I have collected here some ideas about the relationship between text and image in the digital, and I have suggested that with the advent of stochastic ML in the form of artificial neural networks, it is necessary to divide the digital realm into sequential and connectionist subfields. Further, I have argued that only in the digital realm can be found what one might call operative ekphrasis: there, texts do not represent images, but perform them by computationally effecting them. And corresponding to the connectionist and sequential approaches, there seem to be two distinct types of operative ekphrases, involving two distinct notions of language—one emphasizing a pragmatic, another a semantic dimension; both of which, to reiterate, are very much below the full meaning of these words, but with some reasonable connection to them nevertheless. However, against the orthodoxy of computers as only having syntax without semantics, there is at least the possibility that multimodal AI, in its conceptual neurons, in fact encodes meaning—a type of artificial semantics that does not mean to the full extent in which humans mean, but means nonetheless.

The argument I have put forward, then, has both a concrete and a methodical dimension. On the one hand, it serves as an aesthetic analysis of AI that takes into account the technical substrate of its media. What this amounts to is a case for multimodality in discussing these works. It shows that 'there are no visual media', as was said by Mitchell, for whom the separation of mediums always ignores the entanglement of the senses and the linguistic basis of their transmission.[65] At the same time, we have neither Lessing nor Horace to follow, but something else that goes beyond these options. On the other hand, however, this argument was also an example of how Critical AI Studies might not only think about or against, but also *with* AI. My proposed term, operative ekphrasis, was in this case less meant to add a new dimension to an old and venerable concept. Rather, it served as way of thinking about a problem that puts it into a specific situation to see how it fares; in this case, the problem to be studied was the connection of text and image, and the interaction between a technical metaphorics and its humanist use.

These are interesting times—on a technical level, progress in AI is in hyperspeed, and a little more than four years ago, computer-generated grammatically correct sentences were remarkable in themselves; now mere descriptions generate images. While we must not get caught up in the AI hype—ascribing to machines characteristics like consciousness or its builders the status of visionaries for whom the rules of fair play no longer hold—we cannot ignore these developments either. If, as pioneers of Critical AI Studies Jonathan Roberge and Michael Castelle write, ML engineers 'see their own behavior in terms of the epistemology of their techniques',[66] then we as humanists may well check our cultural, philosophical, and aesthetic epistemology refracted in current media technologies. Their categories may be slow to catch up with the reality we see in the wild, and yet while scholarship can observe them from a distance or get involved hands-on, it must be open to adjusting its categories. Operative ekphrasis is one such adjustment.

ACKNOWLEDGEMENTS

I thank Jules Pelta Feldman, Catriona MacLeod, Jay David Bolter, Antonio Somaini, and an anonymous reviewer for valuable feedback on various stages of this paper. I am also indebted to the discussants at events at the Electronic Literature Organization Conference 2022 in Como, Italy; the University of California—Berkeley; the University of Illinois—Chicago; the Paris National Institute of Art History; the University of Hamburg; the Technical University Braunschweig; the University of Düsseldorf; and the University of Basel, where I had the opportunity to present versions of this paper.

NOTES

1– See Jonathan Roberge and Michael Castelle, eds, *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies* (Cham: Springer,

2021); *American Literature* 95, no. 2 [special issue on 'Critical AI'] (2023); the newly founded journal *Critical AI* 1, nos 1–2 (2023); and the mailing list 'All Models', http://allmodels.ai, accessed on 3 January 2024.

2– Rita Raley and Jennifer Rhee, 'Critical AI: A Field in Formation', *American Literature* 95, no. 2 (2023): 185–204, at 188.

3– Simon Lindgren, *Critical Theory of AI* (Cambridge: Polity, 2024); Matteo Pasquinelli, *The Eye of the Master: A Social History of Artificial Intelligence* (London: Verso, 2023); Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven and London: Yale University Press, 2021); Wendy Hui Kyong Chun and Alex Barnett, *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition* (Cambridge, MA: MIT Press, 2021); Louise Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others* (Durham: Duke University Press, 2020).

4– Lisa Gitelman, ed., *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013); Adrian Mackenzie, *Machine Learners: Archaeology of a Data Practice* (Cambridge, MA: MIT Press, 2017); Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, *Pattern Discrimination* (Minneapolis: University of Minnesota Press, 2018).

5– Philip E. Agre, 'The Soul Gained and Lost: Artificial Intelligence as a Philosophical Project', *Stanford Humanities Review* 4, no. 2 (1995): 1–19, at 5.

6– Hans Blumenberg, 'An Anthropological Approach to the Contemporary Significance of Rhetoric', in *History, Metaphors, Fables: A Hans Blumenberg Reader*, ed. Hannes Bajohr, Florian Fuchs, and Joe Paul Kroll (Ithaca: Cornell University Press, 2020), 177–209, at 188.

7– Fabian Offert and Thao Phan, 'A Sign that Spells: DALL-E 2, Invisual Images and the Racial Politics of Feature Space', *arXiv*, 26 October 2022, 1–4, at 3, http://arxiv.org/abs/2211.06323, accessed on 12 July 2023.

8– Daniel C. Dennett, *Intuition Pumps and Other Tools for Thinking* (New York: Norton, 2013). According to Dennett, an intuition pump works by providing a simplified example, an analogy or a metaphor that helps make complex concepts more comprehensible and intuitive. It is a way to trigger instincts and intuitions about a situation in the hope of understanding the underlying principles more clearly. In this case, the concrete technology of 'multimodal AI' serves as an intuition pump for the complex concept of ekphrasis. Unlike in Dennett's case, however, I find that not only does the example illustrate the concept, but also it is able to modify it. This approach also informs the collected volume Hannes Bajohr, ed., *Thinking with AI: Machine Learning the Humanities* (London: Open Humanities Press, 2024).

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

10– N. Katherine Hayles, 'Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis', *Poetics Today* 25, no. 1 (2004): 67–90; Hannes Bajohr, 'Algorithmic Empathy: Toward a Critique of Aesthetic AI', *Configurations* 30, no. 2 (2022): 203–31.

11– Jine [i.e., Jasmin Meerhoff], 'They Lay', GitLab, 24 February 2022, https://gitlab.com/nervousdata/they-lay, accessed on 2 May 2022.

12– Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, 'Zero-Shot Text-to-Image Generation', *arXiv*, 26 February 2021, http://arxiv.org/abs/2102.12092, accessed on 12 July 2023.

13– For more on text-to-image models, see the special issue 'Generative Imagery: Towards a "New Paradigm" of Machine Learning-Based Image Production', *IMAGE. The Interdisciplinary Journal of Image Sciences* 31, no. 1 (2023).

14– While the newer third version is integrated into ChatGPT, the interface of DALL·E 2 shown is still available at https://labs.openai.com, accessed on 22 December 2023.

15– Dave Orr, 'Playing with DALL·E 2', Lesswrong, 7 April 2022, https://www.lesswrong.com/posts/r99tazGiLgzqFX7ka/playing-with-dall-e-2, accessed on 2 May 2022.

16– On the issue of mangled hands, see Amanda Wasielewski, 'Midjourney Can't Count', *IMAGE* 37, no. 1 (2023): 71–82, https://doi.org/10.1453/1614-0885-1-2023-15454, accessed on 12 July 2023. For the inability to produce text, see Eliza Strickland, 'DALL-E 2's Failures are the Most Interesting Thing About It', *IEEE Spectrum*, 14 July 2022, https://spectrum.ieee.org/openai-dall-e-2, accessed on 3 January 2024. However, this may be a function of parameter size: Google's Parti model seems able to produce text with a parameter count above twenty billion: 'Parti: Pathways Autoregressive Text-to-Image Model', Google Research, accessed on 12 July 2023, https://sites.research.google/parti/, accessed on 12 July 2023. The same is true for the current version of Midjourney: 'Text in Midjourney V6', Midjourney blog, 22 December 2023, https://mid-journey.ai/text-generation-in-midjourney-v6/, accessed on 22 December 2023. I tried DALL·E 3 as well as GPT-4o with Orr's prompt and still found that the systems output garbled text, although the title is often legible; if one specifies the exact lines of the poem, however, the models can visualise them relatively well now.

17– dall·ery gall·ery, 'The DALL·E 2 Prompt Book, v1.02'. Dall·ery gall·ery: Resources for Creative DALL·E Users, 2022, https://dallery.gallery/wp-content/uploads/2022/07/The-DALL%C2%B7E-2-prompt-book-v1.02.pdf, accessed on 12 July 2023.

18– 'PromptBase', https://promptbase.com, accessed on 12 July 2023.

19– I derive this conceptual distinction from an influential publication that brought neural networks back into fashion under the rubric of 'connectionism': David E. Rumelhart, James McClelland, and Geoffrey Hinton, 'The Appeal of Parallel Distributed Processing', in *Parallel Distributed Processing. Explanations in the Microstructure of Cognition*, Vol. 1: *Foundations*, ed. David E. Rumelhart, James McClelland, and PDP Working Group, 2 vols (Cambridge, MA: MIT Press, 1986), 3–44, at 43. The term 'sequential' for the classic algorithm stems from the same book: David E. Rumelhart and James L. McClelland, 'PDP Models and General Issues in Cognitive Science', in Rumelhart *et al.*, *Parallel Distributed Processing*, 110–46, at 116. For a more extensive discussion, see Bajohr, 'Algorithmic Empathy'.

20– Thomas Haigh and Paul E. Ceruzzi, *A New History of Modern Computing* (Cambridge, MA: MIT Press, 2021).

21– For a non-technical introduction, see John D. Kelleher, *Deep Learning* (Cambridge, MA: MIT Press, 2019). For a more technical discussion, see Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge, MA: MIT Press, 2016).

22– I make this point in more detail in Hannes Bajohr, 'The Gestalt of AI: Beyond the Holism–Atomism Divide', *Interface Critique* 3 (2021): 13–35, https://doi.org/10.11588/ic.2021.3.81304, accessed on 12 July 2023.

23– Fabian Offert, 'Can We Read Neural Networks? Epistemic Implications of Two Historical Computer Science Papers', *American Literature* 95, no. 2 (2023): 423–28, https://doi.org/10.1215/00029831-10575218, accessed on 3 January 2024.

24– Ruth Webb, '*Ekphrasis* Ancient and Modern: The Invention of a Genre', *Word & Image* 15, no. 1 (January 1999): 7–18, at 7.

25– Ulrich Pfisterer, 'Ekphrasis', in *Metzler Lexikon Kunstwissenschaft: Ideen, Methoden, Begriffe*, ed. Ulrich Pfisterer (Stuttgart: J. B. Metzler, 2019), 99–103, at 99; Webb, '*Ekphrasis* Ancient and Modern', 8–9.

26– Webb, '*Ekphrasis* Ancient and Modern', 8.

27– John Hollander, *The Gazer's Spirit: Poems Speaking to Silent Works of Art* (Chicago: University of Chicago Press, 1995), 7.

28– James A. W. Heffernan, 'Ekphrasis: Theory', in *Handbook of Intermediality: Literature–Image–Sound–Music*, ed. Gabriele Rippl (Berlin: de Gruyter, 2015), 35–49, at 35.
29– Leo Spitzer, 'The "Ode on a Grecian Urn" or, Content vs. Metagrammar', in *Essays on English and American Literature*, ed. Anna Hatcher (Princeton: Princeton University Press, 1962), 67–97, at 72.
30– James A. W. Heffernan, *Museum of Words: The Poetics of Ekphrasis from Homer to Ashbery* (Chicago: University of Chicago Press, 1993), 3.
31– Tamar Yacobi, 'Ekphrastic Double Exposure and the Museum Book of Poetry', *Poetics Today* 34, nos 1–2 (2013): 1–52, at 1, 3.
32– W. J. T. Mitchell, *Picture Theory: Essays on Verbal and Visual Representation* (Chicago: University of Chicago Press, 1994), 152–60.
33– Gotthold Ephraim Lessing, *Laocoön: An Essay upon the Limits of Painting and Poetry*, trans. Ellen Frothingham (Mineola: Dover, 2005), chs 15–16.
34– Mitchell, *Picture Theory*, 154.
35– Renate Brosch, 'Ekphrasis in the Digital Age: Responses to Image', *Poetics Today* 39, no. 2 (2018): 225–43, at 227.
36– Espen J. Aarseth, *Cybertext: Perspectives on Ergodic Literature* (Baltimore: Johns Hopkins University Press, 1997).
37– N. Katherine Hayles, *My Mother Was a Computer* (Chicago: University of Chicago Press, 2005), 50, 101.
38– The adjective 'operative', here, is thus meant to be understood quite literally as 'having the character as an operation'. It is not to be confused with Harun Farocki's *operatives Bild*, sometimes translated as 'operative image' or 'operational image', by which he means images used in surveillance and war that do not require linguistic mediation because they act as sensors rather than representations; Harun Farocki, 'Phantom Images', *Public* 29 (2004), https://public.journals.yorku.ca/index.php/public/article/view/30354, accessed on 3 January 2024. While Jussi Parikka, taking up Farocki's idea, highlights the performativity of images themselves, my concept makes text performative insofar as it produces an image; Jussi Parikka, *Operational Images: From the Visual to the Invisual* (Minneapolis: University of Minnesota Press, 2023).
39– Thus Friedrich Kittler could, shortly after having pronounced that there is no software but only hardware, exclude computer graphics from the class of optical media by declaring them essentially alphabetical. A pixel image, he wrote, 'deceives the eye, which is meant to be unable to differentiate between individual pixels, with the illusion or image of an image, while in truth the mass of pixels, because of its thorough addressability, proves to be structured more like a text composed entirely of individual letters'; Friedrich A. Kittler, 'Computer Graphics: A Semi-Technical Introduction', trans. Sara Ogger, *Grey Room* 2, no. 2 (2001): 30–45, at 32. A similar, if historically inverse, identification of text and image is made by Vilém Flusser, who claims that 'The invention of writing is not so much about the invention of new symbols, but about the unfurling of the [two-dimensional] image into [one-dimensional] rows ("lines"); Vilém Flusser, *Lob der Oberflächlichkeit: Für eine Phänomenologie der Medien*, 9 vols (Bensheim: Bollmann, 1993), 1: 67 (present author's translation).
40– Jay David Bolter, 'Ekphrasis, Virtual Reality, and the Future of Writing', in *The Future of the Book*, ed. Geoffrey Nunberg (Berkeley: University of California Press, 1996), 253–72, at 269.
41– Ibid., 270.
42– I refer to the title of Jerome J. McGann, *The Textual Condition* (Princeton: Princeton University Press, 1991), but the idea that everything digital may be understood as text can be found in Jerome J. McGann, *Radiant Textuality: Literature after the World Wide Web* (New York: Palgrave, 2001), 11.
43– 'Virtually', since Yuk Hui suggests that the ontology of the digital is a broad spectrum that runs from 'colorful visual beings' at the interface level to the 'particles and fields' that make up the circuit boards and the electricity running through it; somewhere in the middle, 'at the level of programming', there are 'text files'. For present purposes, I will stick to this middle position; Yuk Hui, *On the Existence of Digital Objects* (Minneapolis: University of Minnesota Press, 2016), 27–28.
44– Loss Pequeño Glazier, 'Code as Language', *Leonardo* 14, no. 5 (2006), http://leoalmanac.org/journal/vol_14/lea_v14_n05-06/lpglazier.asp, accessed on 12 July 2023. For a philosophically more sophisticated version of this argument, see Juan Luis Gastaldi, 'Why Can Computers Understand Natural Language? The Structuralist Image of Language Behind Word Embeddings', *Philosophy & Technology* 34, no. 1 (2021): 149–214.
45– Stevan Harnad, 'The Symbol Grounding Problem', *Physica D: Nonlinear Phenomena* 42, no. 1–3 (1990): 335–46.
46– Florian Cramer, 'Language', in *Software Studies: A Lexicon*, ed. Matthew Fuller (Cambridge, MA: MIT Press, 2008), 168–74, at 168–69.
47– However, there is a lively debate about how useful it is to speak of a pragmatics of programming languages in a broader sense. One suggestion is to say that 'The pragmatic effects of the program in execution [...] cause changes to occur in the internal state of the computer'; J. H. Connolly and D. J. Cooke, 'The Pragmatics of Programming Languages', *Semiotica* 151 (2004): 149–61, at 154. Benjamin Bratton likewise suggests that 'code is a kind of language that is executable. [...] In this sense, linguistic "function" refers not only to symbol manipulation competency, but also to the real-world functions and effects of executed code'; Benjamin Bratton and Blaise Agüera y Arcas, 'The Model Is the Message', *Noema*, 12 July 2022, https://www.noemamag.com/the-model-is-the-message, accessed on 2 August 2022.
48– Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, 'Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions', *arXiv*, 20 February 2023, http://arxiv.org/abs/2209.03430, accessed on 12 July 2023; Cem Akkus, Luyang Chu, Vladana Djakovic, Steffen Jauch-Walser, Philipp Koch, Giacomo Loss, Christopher Marquardt, Marco Moldovan, Nadja Sauter, Maximilian Schneider, Rickmer Schulte, Karol Urbanczyk, Jann Goschenhofer, Christian Heumann, Rasmus Hvingelby, Daniel Schalk, and Matthias Aßenmacher, 'Multimodal Deep Learning', *arXiv*, 12 January 2023), http://arxiv.org/abs/2301.04856, accessed on 12 July 2023.
49– For this distinction, see Chip Huyen, 'Multimodality and Large Multimodal Models (LMMs)', *Chip Huyen blog*, 10 October 2023, https://huyenchip.com/2023/10/10/multimodal.html, accessed on 3 January 2024.
50– OpenAI *et al.*, 'GPT-4 Technical Report', *arXiv*, 18 December 2023, https://doi.org/10.48550/arXiv.2303.08774; and OpenAI, 'GPT-4V(ision) System Card', 2023, https://openai.com/research/gpt-4v-system-card, accessed on 3 January 2024.
51– Google Gemini Team, 'Gemini: A Family of Highly Capable Multimodal Models', 2023, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf, accessed on 16 October 2023; OpenAI, 'Hello GPT-4o', 2024, https://openai.com/index/hello-gpt-4o, accessed on 23 May, 2024.
52– See Bajohr, 'Algorithmic Empathy' for an attempt to make this medi(a/um) specificity aesthetically fruitful for aesthetic AI; in a way, the present study offers a counterargument to this earlier article by highlighting not the separation but the collapse of text and image medi(a/ums). At the same time, it confirms the insight that any discussion of aesthetic AI needs to be aware of its technical substrate.
53– Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, 'Attention Is All You Need', in *Advances in Neural Information Processing*

Systems, ed. Isabelle Guyon, Ulrike Von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S. V.N. Vishwanathan, and Roman Garnett (Red Hook, New York: Curran Associates, 2017), 5998–6008. For a step-by-step explanation, see also Jay Allamar, 'The Illustrated Transformer', 2018, https://jalammar.github.io/illustrated-transformer, accessed on 12 July 2023.

54– For diffusion models, see Prafulla Dhariwal and Alex Nichol, 'Diffusion Models Beat GANs on Image Synthesis', *arXiv*, 1 June 2021, https://doi.org/10.48550/arXiv.2105.05233, accessed on 12 July 2023.

55– Ryan O'Connor, 'How DALL-E 2 Actually Works', *Assembly AI*, 19 April 2022, https://www.assemblyai.com/blog/how-dall-e-2-actually-works/, accessed on 12 July 2023.

56– In recent memory, Emily Bender and Alexander Koller were the most influential theorists to argue that large language models like ChatGPT—which are unimodal—are not able to operate with meaning; Emily M. Bender and Alexander Koller, 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* [Online] (Association for Computational Linguistics, 2020), 5185–98, https://aclanthology.org/2020.acl-main.463/x accessed on 12 July 2023. However, because they understand meaning as text being 'grounded' in the world, they have to allow for the possibility that multimodality may lead to a model learning 'some aspects of meaning', because it grounds text data in image date; ibid., 5193. I call this phenomenon 'dumb' meaning and explain it in more detail in Hannes Bajohr, 'Dumb Meaning: Machine Learning and Artificial Semantics', *IMAGE* 37, no. 1 (2023): 58–70, https://doi.org/10.1453/1614-0885-1-2023-15452, accessed on 12 July 2023.

57– Charles G. Gross, 'Genealogy of the "Grandmother Cell"', *The Neuroscientist* 8, no. 5 (2002): 512–18, https://doi.org/10.1177/107385802237175, accessed on 12 July 2023.

58– Rodrigo Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried, 'Invariant Visual Representation by Single Neurons in the Human Brain', *Nature* 435, no. 7045 (2005): 1102–07, at 1102, 1106, https://doi.org/10.1038/nature03687, accessed on 12 July 2023. A 2009 study by some of the same authors explicitly call this phenomenon multimodal; Rodrigo Quian Quiroga *et al.*, 'Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain', *Current Biology* 19, no. 15 (2009): 1308–13, https://doi.org/10.1016/j.cub.2009.06.060, accessed on 12 July 2023.

59– Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah, 'Multimodal Neurons in Artificial Neural Networks', *Distill* 6, no. 3 (2021), https://doi.org/10.23915/distill.00030, accessed on 12 July 2023. A more recent paper, now using a diffusion model (similar to GLIDE), arrives at similar results; Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao, 'Cones: Concept Neurons in Diffusion Models for Customized Generation', *arXiv*, 9 March 2023, http://arxiv.org/abs/2303.05125, accessed on 12 July 2023.

60– However, while most researchers agree that there are mostly likely no *single* grandmother neurons (one neuron for one concept), there is a relatively broad consensus that, over time, concepts are indeed stored as sparse rather than dense neural encodings in the brain; Marcel Bausch, Johannes Niediek, Thomas P. Reber, Sina Mackay, Jan Boström, Christian E. Elger, and Florian Mormann, 'Concept Neurons in the Human Medial Temporal Lobe Flexibly Represent Abstract Relations between Concepts', *Nature Communications* 12, no. 1 (2021): 6164, https://doi.org/10.1038/s41467-021-26327-3, accessed on 12 July 2023. What this means is that while the brain may store many individual features for single entities ('dense encoding'), in the long run, such encodings merge along shared features covering groups of entities ('sparse encoding') located in smaller neuronal clusters. This process of grouping more entities under a shared encoding is understood as 'concept learning'. Moreover, a recent study explicitly found that this process of concept learning—from dense to sparse encoding—is similar to and can be simulated in artificial neural networks, suggesting, if not homology, at least analogous behaviour; Louis Kang and Taro Toyoizumi, 'Distinguishing Examples While Building Concepts in Hippocampal and Artificial Networks', *Nature Communications* 15, no. 1 (2024): 647, https://doi.org/10.1038/s41467-024-44877-0, accessed on 3 January 2024. An OpenAI study likewise found that neural networks use a sparse encoding for concepts; Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu, 'Scaling and Evaluating Sparse Autoencoders', arXiv, 6 June 2024, http://arxiv.org/abs/2406.04093, accessed on 24 June, 2024.

61– Bender and Koller, 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data', 5185–98.

62– Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992).

63– As such, multimodal AI is more than remediation, as Jay David Bolter suggested, since this term still keeps the separation of media intact; Jay David Bolter, 'AI Generative Art as Algorithmic Remediation', *IMAGE* 37, no. 1 (2023): 195–207.

64– Liliane Louvel, *Le tiers pictural* (Rennes: Presses Universitaires de Rennes, 2010).

65– W. J. T. Mitchell, 'There are No Visual Media', *Journal of Visual Culture* 4, no. 2 (2005): 257–66, https://doi.org/10.1177/1470412905054673, accessed on 12 July 2023.

66– Jonathan Roberge and Michael Castelle, 'Toward an End-to-End Sociology of 21st-Century Machine Learning', in *The Cultural Life of Machine Learning*, 1–30, at 6.