Studi di estetica, anno LII, IV serie, 3/2024
ISSN 0585-4733, ISSN digitale 1825-8646, DOI 10.7413/1825864695

Charlotte Bolwin

Digital *ekphrasis*?
On language-image-relations in contemporary AI's imageries

Abstract
This essay takes the concept of "digital ekphrasis*" as an opportunity to look at contemporary multimodal AI – or more precisely text-to-image generators, understood as the latest phenomena in the media history of technical images. In my discussion, I raise the question of whether the digitally programmed image generation performed by programs like Stable Diffusion, Midjourney or DALL-E can be thought of as ekphrasis. Following recent discussions in the field of media theory, I thereby ask whether the criterion of operativity is decisive for distinguishing text-to-image generation from ekphrasis in the classical sense. My discussion evolves in a revision of ekphrasis in the art historical sense, confronted with the structural processes of multimodal AI. However, the comparison of these two modes of* ekphrasis *reveals how impoverished the concepts of both image and language risk being in the context of text-to-image-modelling. This, in turn, does not mean that current AI imageries cannot be discussed regarding a (post-) digital aesthetics. Rather, as recent media artists working with multimodal AI show, text-to-image reformulates old questions about the relationship between art or aesthetics and (media) technology.*

Keywords
Image theory, Text-to-Image, Postdigital aesthetics

Received: 19/06/2024
Approved: 25/11/2024
Editing by: Lorenzo Manera

charlotte.bolwin@uni-weimar.de (Bauhaus-Universität Weimar)

Image generated with *Stable Diffusion*'s 2024 version

1. *Computers, images, and the notion of an "operative* ekphrasis*"*

Since its invention in the 20th century, the computer has become an omnipresent medium and an integral part of our current technological environment. Today, the universal and discrete "writing-reckoning machine" (Kittler 1996: 245)[1] is the basis for a vast multitude of operations, so comprehensive in scale that it no longer appears reasonable to speak of singular acts, scenes and settings of digitalization, but rather of *digitality* – an encompassing techno-culture in which our ways of acting, knowing and perceiving are consistently organized alongside or within digital media (among others, see Stalder 2016; Distelmeyer 2021). Within this culture of digitality, something has become evident that 20th century media studies had not foreseen: the 21st century marks the establishment of the computer as an *image medium*. As German Sybille Krämer (1988) has pointed out, computers are mathematically structured "symbolic machines", meaning that they operate by means of a dynamic alphanumerical code that can assume various forms of actualization but never leave its basic formal structure. And yet, even though they are calculating machines, computers have become visual media in the sense that they are

[1] For this term (which was originally formulated in German as "Schreib-Rechen-Maschine"), cf. Kittler (1996: 245).

used for processing, that is, altering and generating images[2]. This is illustrated not least by the various methods of image generation that are currently available, ranging from computer graphics in the classical sense to the mundane phenomena of algorithmically optimized smartphone photography and to text-to-image engines. The latter allow for visual phenomena – images for now, but video is to follow[3] – to be generated through predefined semantic features from textual descriptions, the so-called *prompts*. In view of the similarity to the images drawn, painted, photographed, or designed by humans, this computer-generated imagery (CGI) is mostly not even recognized as computer-generated at all[4]. Hardly distinguishable from human-made images, CGI is currently being used primarily in marketing and web design. They have also triggered debates regarding the questions of authorship, originality, and creativity in the field of art[5].

In simple terms, CGI is just another form of technical imagery, with the caveat that "technical" relies on forms of datafication and cloud or networked computing. The question of the novelty or difference of such technical images as compared to their predecessors is commonly answered in media philosophy with an emphasis on the structural features of these images and the possible new ontologies derived from them. However, computer graphics can also be placed in a continuous media history of technically generated images, as Vilém Flusser pointed out as early as 1985, with roots both technically and aesthetically in video, TV, film, and analog photography (cf. Flusser 1985). By contrast, one of Wal-

[2] Regarding these three functionalities of media, see Winkler 2015.

[3] As *The Guardian* reports in spring 2024, *OpenAI* is to launch a text-to-video-software under the name Sora, and this is most likely just the beginning of a vast range of software that can be used to generate moving images. Available at: https://www.theguardian.com/technology/2024/feb/15/openai-sora-ai-model-video [last accessed 26/02/2024].

[4] Even though there are much popular discussions about bad AI images whose errors conspicuously point to the involvement of a machine, there are numerous examples of how indistinguishable AI-generated images have become from images in whose creation humans were physically and situationally involved. The issue of so-called deep fakes not only concerns users of social media platforms but has also become a focal point in contemporary art. For example, Roop Rainisto explores this increasingly nebulous boundaries between synthetic images and photography in his artworks.

[5] In this context, the debates about so-called AI art, which have recently been stimulated by the fact that images generated with programs such as Midjourney have received art awards – and thus crucial recognition as elements and results of artistic practice or even as works of art. Examples of this are text-to-image-based images such as *Théâtre d'Opéra Spatial*, created by Jason M. Allen, or the *Portrait of Edmond De Belamy*, created by the art collective *Obvious*.

ter Benjamin's main assumptions regarding modern media such as photography was that these bring genuinely new aspects of worldly phenomena to light – Benjamin's concept of the optical unconscious being just one example (see Benjamin 1939) – and have the potential of opening and establishing new collective modes of perception. In other words, the question remains how the nature of a medium changes the reality displayed in it – a question that reiterates the fundamental idea of media-aesthetic thought that perceived reality can never be immediate.

Accordingly, media theoretical discourses on the image are often less concerned with a *qualitative* distinction between automatically and manually produced images such as paintings or drawings – the question rather being how images derived from black-boxed and automated production imply a new cultural paradigm in which established relations between objects, media and viewers are (re)arranged during technological operations. A paradigmatic example for this is photography, which has produced a shift in the philosophical discourse on images from the relationship between image and reality toward a novel factual-material connection between the depicted object and its medium of representation. While painting is seen as producing images, photographs simultaneously produce documents, hence factual, material "emanations" of the real, as Roland Barthes pointed out (1980: 126). The indexical value of the image for which photography stands like no other medium, was in turn contested by *digital* imagery in the 20th and 21st century – from early digitized photographs to the latest computer-generated images that are constructed and rendered entirely within the discrete calculating system of the computer.

While digital imagery has played a crucial role in redirecting the discourse of media theory toward the connection between images and a given reality, the current rise of text-to-image engines prompts us to revisit the question of the relationship between images and texts or between visuality and language. In strictly theoretical terms, this relationship involves two media forms in which information can be presented and perceived; information which by nature of the computer *always* has a binary structure. It must be characterized as a relationship that is itself not immediate but mediated, as image and text never encounter each other outside of their translation into the binary language of computation. Still, AI-driven text-to-image modeling is a scene of media production that is

based on an existent connection between language and image[6]. Since text-to-image models such as *Midjourney*, *DALL-E* or *Stable Diffusion* use linguistic inputs to generate images[7], it is thus tempting to think of text-to-image modeling as a form of digital *ekphrasis*. After all, what else could it be than the language-based evocation of images, which (just) happens to be embedded in digital technology? It is this problem of superficiality that German philosopher and literary scholar Hannes Bajohr (2024) aims to tackle by introducing the term "operative *ekphrasis*". Clearly not intended to describe *all* digital images, but distinctively the named text-to-image phenomena produced by AI, operative *ekphrasis* according to Bajohr (2023: 77) is based on the "collapse of the text/image distinction" in the process of text-to-image modelling (s.o.) or more precisely in the fact that "this technology can process both text and image as *one* type of data"[8]. In this sense, as Bajohr argues, "multimodal AI does away with the separation of mediums (*sic*) that is at the core of *ekphrasis*" to instantiate a novel form of "operative *ekphrasis*" based on what he identifies as "performative" digital code (2023: 77)[9].

In the following, I will reassess the connection between the art-historical concept of *ekphrasis* and image synthesis in contemporary, multimodal AI[10]. In doing so, I pursue two main concerns: Firstly, acknowledging the fact that Bajohr draws on theoretical discourses on the mediality of the

[6] In this essay, the acronym AI, which is an abbreviation for the term artificial intelligence, is used in a heuristic-pragmatic sense. In the context of the media-theoretical and philosophical debate on so-called artificial intelligence, it has rightly been pointed out that the term can be understood as part of an anthropomorphizing and naturalizing terminology that is often used affirmatively – for example in the economic and political marketing of machine learning processes. At the same time, the term artificial intelligence has been associated with such technologies since the early days of machine learning. For example, it already appears in the context of the Dartmouth Conference.

[7] Reflecting the author's impression that the functionality of text-to-image engines becomes clear at first sight when approaching the interface, see: https://stablediffusionweb.com/ or https://openai.com/dall-e-3 (accessed 26/02/2024).

[8] In fact, this view can be challenged, because designers do not use the same model to process images or text, which shows that they are not one type of data in the strict sense: With images, you process 2D spatial relationships between image areas; with text, you process linear sequences of tokens. A translation, a bridge between text and image (or use your language from below: equivalence / correspondence) is therefore required, and the labeling is this bridge.

[9] It should be noted here that this only applies if we (like Bajohr) assume that the medium is the computer in general. On the other hand, if you specify that different models are different media, the quoted view becomes debatable.

digital, which place the concept of operativity at their core, I will take a critical approach to the definition of "digital *ekphrasis*"(cf. Viglialoro, Latini 2024) as operative, reflecting that it is already possible to speak of operativity in the classical concept of *ekphrasis* in terms of the performance of language and image. In this light, it appears questionable whether *operativity* can be upheld as the categorial difference between aesthetic ekphrastic relations and algorithmic semantizations. This, in turn, raises suspicion about the idea that the key to understanding the phenomenon of AI-based text-to-image modeling rather is to be found in the differences between an aesthetic and a computational plot. Therefore, secondly, I (want to) argue that the image-text relations in multimodal AI are not necessarily to be determined by means of a novel or different characterization of the relationship between language and image, but rather via the reductive use of both media – language *and* text – that appears necessary for AI-based image generation processes to untangle their mutually complex medial performance. This becomes particularly evident in the technical operation of *labeling*, in which any textual or visual expression is reduced to a quasi-empty signifier, signifying not aesthetic value, but a reductive attribution and combination of the possible deriving distinct features.

2. *Inter/medialities: revisiting the ekphrastic passage*

The avid debate around the concept of *ekphrasis* that pervades numerous disciplines including philosophical aesthetics and art history, image theory or literature and religious studies, to name just a few, has fueled a vast theoretical discourse on the intermedial relations between the visual (or, at times, the sensual) and language. Since this discourse is distributed not only across disciplines but also across epochs and across objects, it can be hard to keep track of a distinct definition. One instance of this confusing situation is the fact that the lemma "ekphrasis" does not have its own entry in the German handbook of aesthetics – *Ästhetische Grundbegriffe* (Barck *et al.* 2001) – which is one of the most pertinent works of reference for the fundamental concepts of art theory and history for German-speaking scholars in the field of art history and theory. On the other hand, there are countless anthologies dedicated to the topic of *ekphrasis* as a historical practice or within the scope of image theory, two of which are particularly significant: the volume *Beschreibungskunst – Kunstbeschreibung. Ekphrasis von der Antike bis zur Gegenwart*, edited by

Gottfried Boehm (1995), and W.J.T. Mitchell's *Picture theory: essays on verbal and visual representation* (1994).

Despite the vastness of the scholarly discourse, a key distinction can be made: while *ekphrasis* in the narrower, art-historical sense is often used synonymously with the description of images, or rather artworks[11], *ekphrasis* initially meant the possibility or practice of evoking images by means of verbal, often poetic expression. The *images* produced thereby were originally *mental* images arising in the listeners as they immersed themselves into the qualities of the object described – be it an image, a material item, a living body, or a natural site. Perhaps the earliest example of an *ekphrasis* is the description of Achilles' shield in Homer's *Iliad* (see Mitchell 1994: 152). It not only concerns the appearance and sensual qualities of the legendary shield, but also tells the story of its creation. Initially rooted in ancient aesthetics and directly connected to its epic, rhetorical and poetic artifice, the idea as well as the method of *ekphrasis* in the 18th and 19th centuries change from an art of description to a description of art (see Rosenberg 1995: 302). As Raphael Rosenberg observes (this transition), *ekphrasis* throughout the 18th and 19th century shifts from an independent aesthetic form to a genre that could soon be called an "auxiliary science" of art history. It is, unsurprisingly, this period of early modernity that challenges language to submit to the criteria of objectifiability. As Rosenberg underlines, in the context of a broader movement toward scientific objectivity, descriptions are no longer meant to display a poetics of their own, and in ekphrastic accounts of artworks the intensity and immediacy of the aesthetically pure impression yields to sobriety and objectivity (see Rosenberg: 314, 316). In the 20th century, according to Rosenberg, the evocative, "expressionist" style which employs a poetic use of language returns to the scene to complement the "analytical", rather technical style – in situations that allowed for vagueness (Rosenberg: 317).

Departing from this narrow definition of *ekphrasis* as a describing practice that originates in art history, the term can also be understood – in a much broader sense – as the relation of language to visual culture *in toto*. It is in this spirit that Jaś Elsner (2010) wrote his essay *Art history as* ekphrasis. Instead of examining *ekphrasis* within the scope of art history, Elsner argues for an understanding of the entirety of art history as an ekphrastic discourse: "art history […] is nothing other than *ekphrasis*", he

[11] Already in 1940, Erwin Panofsky noted that the term "ekphrasis" was then used almost exclusively in the sense of art description. See Panofsky 1940: 20-1.

writes, "or more precisely an extended argument built on *ekphrasis*", as "it represents the tendentious application of rhetorical description to the work of art […] for the purpose of making an argument of some kind to suit the author's prior intent" (Elsner 2010: 11). From this perspective, the form and style of *ekphrasis* as well as the choice of its object may have been influenced by external motives all along – an idea which attenuates the notion of *ekphrasis* as a "spoken image".

But not only does *ekphrasis* establish what Elsner calls "the descriptive basis for the practice of art history" (Elsner 2010: 12); more broadly, it defines the intermedial relation between language and the sensual phenomena of the world. In this sense, *ekphrasis* is a performative and thoroughly *aesthetic* procedure. The "descriptive act" of *ekphrasis* is not just instrumental, nor is it completely absorbed in functionality – it is rather an operation that organizes "a movement from [image] to text, from visual to verbal" (Elsner 2010: 12). Thus, an irresolvable but productive tension between aesthetic modes of perception and medial forms of communication is set to work: "Not everything in the world of the sensual autonomy of the object can be translated into words, and much that was not there is inevitably added by words" (Elsner 2010: 12). Moreover, Elsner notes the dependence on verbalization and description that is inherent specifically to the intersubjective configurations of art: "images and objects – insofar as they are designed to relate to us all – *invite ekphrasis*, indeed they require it" (Elsner 2010: 13).

Although Elsner argues in favor of an intradisciplinary understanding of *ekphrasis* as a means of art history, the fundamental performative and aesthetic character of ekphrastic relations come to the fore in his considerations, as *ekphrasis* is not only necessary to describe and historicize art, but "is inevitable in the viewing" and experience of artworks as sensual artifacts (Elsner 2010: 13). In this reading of Elsner, *ekphrasis* is a "mode of understanding" guided by interpretative or appropriative interest which thus *has* a functional character, but also opens what could be called an "*ekphrastic* passage": a space of aesthetic permeability that is determined by the tension between the modalities of word and image, where what is unspeakable to the speaker can be imagined by the listener and what is invisible becomes conceivable in words. It is, writes Elsner, an "attempt to make the object speak" (Elsner 2010: 25), by way of using language to make it appear and perform as an image. The "resistance [of the object] to being fully verbalized" (Elsner 2010: 26) is counteracted by the stubbornness of language that seeks to actualize a form, a linguistic picture, that is rhetorically appealing by its own standards.

By revisiting the ekphrastic passage I aim to clarify two important aspects pertaining to the question of a digital *ekphrasis*. On the one hand it becomes apparent that the ekphrastic passage has been operative long before the advent of AI imagery; it has always performed an intermedial *operation* – or rather different, folded and recursively structured operations of translating and reverse-translating, actions of relating, of reproducing the perceivable and thereby creating and changing the perceptible. Whether as a rhetorical-literary practice or as a (supposedly) objective description of artworks and images, *ekphrasis* is embedded in a performativity of language that responds to the ambivalences of the image. And, on the other hand, it is precisely this tension, organized and staged through ekphrastic passages, that gives rise to the aesthetic experience through which *sense* is organized in its twofold meaning of sensual and intelligible[12]. Consequently, neither language nor image can be deemed simple means to an end but generate an irreducible aesthetic complexity.

As philosopher and art historian Gottfried Boehm puts it, *ekphrasis* is essentially a passage between the potentials of two media – and hence a scene of their inner, already-given relatedness: "The silent image and the blind language each had their complement in the other medium", he writes (Boehm 1995: 9). In this sense, the ekphrastic passage is not a superficial or merely formal image-text relation, but refers to a deeper, synesthetic tension that is reducible neither to a strict opposition of two media nor to a linear translation between them. Rather, *ekphrasis* exhibits a non-equivalent relation between the potential linguistic quality of the image and the potential visuality of language. In what follows, I argue that this dynamism is depleted in the programmed imaginaries of text-to-image models, insofar as these models enact one-dimensional and formalistic concepts of image and language. To understand this, we must take a closer look at the technical systems that stage text-to-image processes in their currently known forms.

[12] For this double meaning of "sens", derived from the French *sens*, see especially the aesthetic theory of French philosopher Jacques Rancière (2000).

3. *Multimodal AI and the mechanism of labeling*

At the beginning of this text, there is an image. In bright colors, it extends across the entire width of the page in a stretched landscape format. Showing lightly and colorfully dressed pale women in a palatial building from whose windows the view is lost in the distance, its style may suggest at a first glance that we are looking at a baroque painting. As the image caption indicates, however, this image is not a scan or photograph of a historical oil painting, but a digital image-file generated with *Stable Diffusion*, a multimodal AI or text-to-image-program based on a collaboratively developed open-source software[13]. The corresponding image was generated with the linguistic input "create a scene of 'ekphrasis'" combined with the predefined choice "in the style of the Baroque"[14]. This *prompt* is the textual request that constitutes the interaction between the human user and the machine, which then translates the given input value into a visual output in the form of a digital image within seconds[15]. To simplify (and standardize) such a computer-aided form of image generation, systems like *Stable Diffusion* or also *Midjourney* and *DALL-E* work with a graphic user interface that enables textual input – an interface similar to the ones used for search engines, messaging services, or chat bots, or also users use when using a search engine or chatting with friends or service bots, or also the one displayed when interacting with the "smart intelligence" of ChatGPT.

Like other text-to-image models, *Stable Diffusion* is based on a stochastic machine-learning structure colloquially addressed as artificial intelligence or "AI"[16]. In critical AI studies, "AI" refers to a set of processes engaging machine learning "as an assemblage of technological arrangements and sociotechnical practices, as a concept, ideology, and *dispositif*" (Raley, Rhee 2023: 188). "Multimodal AI" on the other hand designates structures that are designed to process multiple medial forms like texts,

[13] For more information on Stable Diffusion's set up, see: https://stability.ai/news/stable-diffusion-public-release (accessed 26/02/2024).

[14] I generated the image myself in February 2024 with Stable Diffusion's online interface in version 2. In editing this text, version 3 has been announced: https://stability.ai/news/stable-diffusion-3 (accessed 26/02/2024).

[15] It is to be noted that for instance *Midjourney* in its latest versions also supports visual prompts, i.e. image-to-image modeling.

[16] Here I am partly relying on a lecture from the Weimar lecture series "Feeds and Flows" on November 27th, 2023, given by computer scientist Niklas Decker under the title *From noise to art: user-controlled image generation beyond prompt engineering*.

sounds, and images. Without delving into the broad field of debates on machine learning and artificial intelligence or providing a detailed explanation of how the so-called neuronal networks which multimodal AI is based upon function, I here wish to question the speculative term of digital *ekphrasis* regarding its pertinency and adequateness when it comes to AI-based imagery as the latest form in the history of technical images. My focus is on the concepts of language and image put to work in these models, because, as explained above, *ekphrasis* is in fact concerned with an aesthetic, performative understanding of both.

The fact that the generation of images is based on linguistic prompts suggests a structural analogy to the concept of *ekphrasis*, leading Bajohr to assume that AI-generated images are "ekphrastic […]: producing visual constellations through text" (Bajohr 2024: 83). The need of the input text to be as "descriptive" as possible – furthering Bajohr's argument – is highlighted by the emergence of a so-called promptology, i.e. the online communities' lore of achieving the highest possible precision in the input in order for the output to come as close as possible to the user's idea[17]. On closer inspection however, this aspect in fact points us to a characteristic of the underlying symbolic computing processes that *Stable Diffusion*, *Midjourney* or *DALL-E* operate with, which renders such assumptions questionable. After all, the aim of precision clashes with the basic way in which the image generator makes randomized choices from a vast number of possibilities in a matter of seconds. Therefore, rather than being evocative or descriptive to the extent of describing the smallest of details, the lore of promptology resorts to the hack of requesting a certain style or genre which the desired details are associated with. This is reflected in the interface of *Stable Diffusion* offering predefined input terms – also marked in the interface as buttons – such as "in the style of the Baroque", which channels the input to certain resulting forms and color schemes. The same mechanism of stochastic machines learning requires that any form of ambivalence or vagueness (on which both language and image [happen to] thrive) be eliminated in the computing process to generate an image that can be seen as matching the requirements.

[17] As a field of practice, promptology describes the intentional specification of prompts by human users to operate the image generation machine in the best possible way. The aim can be a certain ideal result as well as the crossing of the algorithmic logic by glitches or other disturbances of the digital system. In keeping with the growing desire for a design that is as free and differentiated as possible, Stable Diffusion provides a "prompt database" for its users, see: https://stablediffusionweb.com/de/prompts (accessed 2/4/2024).

Images, as Kate Crawford and Trevor Paglen (2021: 1107) put it, being "remarkably slippery things, laden with multiple potential meanings, irresolvable questions, and contradictions", pose a challenge to AI that it responds to by means of "labeling" (or "data labeling") processes. Labeling in the context of machine learning assigns exact correspondences between all kinds of linguistic and visual signifiers and the statistically relevant signified. In other words, by evaluating text-image relations from an enormous quantity of data, the machine establishes a "semantic" system (see Bajohr 2024: 77) of its own, which must remain purely one-dimensional to meet the requirements of the algorithmic structure.

According to Katia Schwerzmann and Alexander Campolo (2023), labeling constitutes one of the basic operations characteristic of machine learning as an example-based programming paradigm. In the context of image generation, the objective for AI is to produce a final result that matches given requirements: "In contrast to the rule-based programming logic, where rules are prescribed explicitly and abstractly in advance, the example-based logic of machine learning begins at the end: with a set of desired concrete outputs" (Schwerzmann, Campolo 2023: 5). The creation of exact "semantic" correspondences between the signifier and the signified, therefore, is not only crucial for the success of the training process, it also is an inevitable effect of the specific logic of the model. As Schwerzmann and Campolo further explain, labeling not only means unambiguously linking linguistic units (so-called categories) with other instances—in our example, visual signs. The labeling of data, by which generative models are trained, is at the same time a practice of reduction in which ambiguity, ambivalence and diversity are eliminated so that the model learns a probability distribution that can be extrapolated to previously unseen images: "this well-understood process involves subtleties that point not only to a single normative moment of ascription of labels but also to a host of other practices that permit the emergence of norms through aggregation, making certain features intelligible while discarding others" (Schwerzmann, Campolo 2023: 5).

What follows from this is not only that any ambiguities or unexpected 'creative' effects ostensibly produced by AI cannot be a result of complex semantic or genuinely *aesthetic* relations. What is more, these effects derive from the randomness with which the algorithm makes choices from a seemingly endless number of possibilities, and the difficulty we ourselves experience with handling such contingency. Moreover, as Crawford and Paglen demonstrate, AI's training process requires an allegedly limitless amount of data from which it derives its semantic system by

statistical means. “Images do not describe themselves” (Crawford, Paglen 2021: 1107) – in order to ‘learn’ about the meaning of a given image, the algorithm depends on finding some piece of textual information previously associated with it. As the data that is available to web crawling in many cases provide text-image relations that are one-dimensional at most, this affects not only the semantic depth of the AI’s understanding but also makes it susceptible to nonsensical or politically problematic associations.

If, considering the above, we understand the linking of text and image in text-to-image models as a functional assignment of data or data sets – as a “process of matching single instances [...] with a generic label” (Schwerzmann, Campolo 2023: 6) – then it becomes clear how reductive and normative the foundations of this computerized performance or creativity ultimately are. As a result, the images created by these processes are not only generated, but also generic – and the same applies to the form of linguistic description on which their operative *ekphrasis* is based. It therefore becomes clear that text-to-image models cannot be based on the intractable mediality of language and image, let alone the intermedial relation that constitutes what has been described as theekphrastic passage. For, as scholars have long spelled out, the ekphrastic relation is ultimately based on an aesthetic understanding of language and images. It is rooted in their performativity and their polysemy as well as in the potentials of an aesthetic synthesis. My argument at this point is not that the structural comparison between text-to-image and the image description of *ekphrasis* is absurd or completely wrong. However, the idea that we are still dealing with an ekphrastic paradigm in text-to-image models ultimately leads to a radical depletion of the categories of image and language and, consequently, to a degradation of *ekphrasis* as a concept.

4. *AI imagery and the perspective of post-digital aesthetics*

To answer the question of whether multimodal AI can be thought of as a form of digital *ekphrasis*, it is essential to specify how the ekphrastic passage is understood: as an aesthetic relation between language and imagery, or – as digital technology suggests – as a formal, merely structural, and thus codifiable relation in a system of clear values and references. This essay has suggested that there is a considerable gap between aesthetic *ekphrasis* that is based on the rhetorical and poetic potentials of both (human) language and of images, which it uses to appeal to the image-relatedness of our imagination, and digital image synthesis through

multimodal AI. This gap not only concerns the relations, differences, or potential convergences (compare Bajohr 2024) between linguistic and visual forms, but ultimately pertains to the aesthetic concepts of word and image itself. While on the one hand we can assume that an aesthetic scene of *ekphrasis* is already operative, the digital setting of multimodal AI seems ekphrastic above all in the fact that language is the apparent basis of its imagery. However, this is not an essential feature of AI-based image generation, but ultimately an interface effect that could as well be reverted to an image-to-text model[18], because it is based on the equivalence of images and labels that the stochastic system learns and then outputs. The relation between the element (*text*) and the element (*image*) remains one-dimensional and ultimately contingent.

It thus appears plausible that text-to-image models primarily depend on a standardizing machinery in which interconnected computers determine which motifs, images and styles correspond to which linguistic terms. Notwithstanding the astounding speed with which new images are constantly being generated, superficially meeting the requirements of the prompt, it becomes more and more obvious that they contain elements that we perceive as deeply redundant and generic. The most pertinent reason for this may be that these models are based on the mimetic paradigm of (photographic or painterly) realism and therefore remain captive to a reductionist, superficial formal language as well as in overt visual language.

It has also become apparent that prompting itself functions as a form of standardization, as the systems neither tend to generate the images that users imagine, nor do they regularly really surprise them. Such structural limitations become – at least at this point – evident for instance when it comes to generating images of abstract concepts, where multimodal AI tends to fail across the board – as can be regarded to be the case in the image used at the beginning of this essay. For the sequence of tokens to be processed by the machine in the best possible way, users must acquire 'promptological' knowledge, removing any doubt that text-to-image engines remain machines. They are production devices, rather than capable independent ekphrastic operators.

Unquestionably, these machines (like all other machines) can be used in artistic and creative ways. Furthermore, it would be overly pessimistic to conclude that the technical images arising from text-to-image models

[18] Compare here the "visual" AI *Astica* which produces textual descriptions on the basis of visual input data such as photographs uploaded into the application's interface: https://astica.ai/ (accessed 26/02/2024).

cannot be part of a techno-aesthetic imaginarium or that they have no aesthetic value whatsoever. Considering text-to-image modeling as a(nother) "cultural technique"[19] that condenses and expands *human* practices through the possibilities of technology, it becomes clear that in inscribing and furthering human realities and imaginations, it can be just as creative or aesthetic as functional or operative. The aim of this essay, however, has been to raise concerns about the idea of conceiving AI image generation as "digital *ekphrasis*" or of speaking of digital *ekphrasis* at all. In the end, *ekphrasis* remains a scientific and aesthetic method that can be used in the *reflection* of digital imageries: namely to describe them critically, and to ask what genuine aesthetic potentials lie in these not necessarily, or not intrinsically aesthetic operations. At this point, it is particularly worthwhile to look at the field of contemporary media artistic practice, where artists pursue both: reflexive exploration and practical appropriation of AI technologies[20]. The tension between both marks (and has marked) the field of media art generally and contours the force field of digital aesthetics.

AI-based image generation should thus not be understood as a process of *ekphrasis* (or as a substitute for it), but as a moment of the technological production of images that perhaps more even than other images brings *ekphrasis* into play anew: as a form of synesthetic reflection that always refers to the technical interfaces preceding the image. Here, the perspective of post-digital aesthetics opens at the interface of human and technological processes; an aesthetics that not only negotiates the forms of *experience* vis-à-vis an increasingly technological world, but also reflexively negotiates the basic terms and concepts of aesthetic theory and practice in relation to technology.

[19] I here refer to the German discourse cultural techniques or "Kulturtechnikforschung" as coined by media theorist Bernhard Siegert (and others following). For a relevant publication outlining this field, see for instance Siegert 2015.

[20] Artists who are currently integrating text-to-image into their practice include Roope Rainisto, Margaret Murphy, Mario Klingemann or Niceaunties and Kevin Abosch. Their works exhibit the alienation of computer-generated imagery, which is experiencing a new boom in the context of recent debates about AI; or the increasingly difficult to recognize/or the increasingly tenuous boundaries between human and technological ways of creating images are explored – for example, by fusing photography with photorealistic AI image material, thus creating what is popularly referred to as 'deep fakes'.

Bibliography

Bajohr, H., *Operative ekphrasis: the collapse of the text/image distinction in multimodal AI*, "Word & image", 2 (2024), pp. 77-90.

Barck., K., *et al*., *Ästhetische Grundbegriffe. Historisches Wörterbuch in sieben Bänden*, Stuttgart, Metzler, 2001.

Barthes, R., *La chambre claire. Notes sur la photographie*, Paris, Gallimard, 1980.

Benjamin, W. [1939], *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit. Drei Studien zur Kunstsoziologie*, Frankfurt a.M., Suhrkamp, 1980.

Boehm, G., Pfotenhauer, H., *Beschreibungskunst – Kunstbeschreibung: Ekphrasis von der Antike bis zur Gegenwart*, München, Brill/Fink, 1995.

Crawford, K., Paglen, T., *Excavating AI: the politics of images in machine learning training sets*, "AI & SOCIETY", 36 (2021), pp. 1105-16.

Distelmeyer, J., *Kritik der Digitalität*, Wiesbaden, Springer, 2021.

Elsner, J., *Art History as Ekphrasis*, "Art history", 33 (2010), pp. 10-27.

Flusser, V., *Ins Universum der technischen Bilder*, Berlin, European Photography, 1985.

Kittler, F., *Computeranalphabetismus*, in J.D. Marcjovski, F. Kittler (Hrsg.), *Literatur im Informationszeitalter*, Frankfurt - New York, Campus Publishing, 1996.

Krämer, S., *Symbolische Maschinen. Die Idee der Formalisierung in geschichtlichem Abriß*, Darmstadt, Wissenschaftliche Buchgesellschaft, 1988.

Mitchell, W.J.T., *Picture theory. Essays on verbal and visual representation*, London, University of Chicago Press, 1994.

Latini, M., Viglialoro L. (eds.), *Digital* ekphrasis, "Studi di Estetica", 1 (2024).

Panofsky, E., *Kunstgeschichte als geisteswissenschaftliche Disziplin* (1940), in Id., *Sinn und Deutung in der bildenden Kunst*, Köln, Dumont, 2002, pp. 7-35.

Raley, R., Rhee, J., *Critical AI: a field in formation*, "American literature", 2 (2023), pp. 185-204.

Rancière, J., *Le Partage du sensible. Esthétique et politique,* Paris, La fabrique, 2000.

Rosenberg, R., *Von der Ekphrasis zur wissenschaftlichen Bildbeschreibung: Vasari, Agucchi, Félibien*, "Zeitschrift für Kunstgeschichte", 58 (1995), pp. 297-318.

Siegert, B., *Cultural techniques: grids, filters, doors, and other articulations of the real*, New York, Fordham University Press, 2015.

Schröter, J., *Medienästhetik, Simulation und Neue Medien*, "Zeitschrift für Medienwissenschaft", 1 (2013), pp. 88-100.

Schwerzmann, K., Campolo, A., *From rules to examples: machine learning's type of authority,* "Big data & society", July-December (2023), pp. 1-13.

Stalder, F., *Digitalität*, Berlin, Suhrkamp, 2016.

Winkler, H., *Prozessieren: Die dritte, vernachlässigte Medienfunktion*, Paderborn, Fink, 2015.