"THIS PROMPT CONTAINS PROHIBITED WORDS": LANGUAGE, EKPHRASIS, AND THE LIMITS OF THE GENERATIVE IMAGINATION

BENJAMIN ZWEIG

ABSTRACT | This essay wonders aloud about the limitations of generative AI imagery in relationship to descriptive language. It looks at the questionably ekphrastic nature of generative AI and asks whether models and platforms such as Dall-E 2 spur creativity or constrain it by the underlying and opaque way the models use literal descriptive language.

KEYWORDS | Generative AI, Ekphrasis, Language, Dall-E 2, Midjourney

This Prompt Contains Prohibited Words

"This prompt contains prohibited words"—this is the admonition that I received after entering what I thought was a straightforward prompt in the AI image generator NightCafe.[1] What had I written to receive this warning and be denied access to my own imagination and the cumulative result of the aggregated power of an algorithm trained upon millions of images? "Massacre of the Innocents."

The biblical story of King Herod ordering the killing of all the male children of Bethlehem, frequently depicted in medieval and early modern art, is not what one might expect to use as a prompt for generative AI. I admit that it was a willfully odd prompt, but the story is well known, and the assumption that a text-to-image generator might "know" the scene and the title from many examples of freely available art that AI models could be trained on was not an unreasonable assumption. To this, "Massacre of the Innocents" provides a succinct caption of the kinds that are used in models like Dall-E 2 and Midjourney. And yet, I was told no. My prompt contained prohibited words.

Words

This was not the first time-generative AI had denied me its creation. Earlier, I had entered a prompt for a female warrior—in the pretty standard vein of Frank Frazetta-type fantasy art—into NightCafe, but was rejected. It took several rounds of prompting to discover that any description of female anatomy beyond generic descriptions like "athletic" would make NightCafe raise the prohibited words flag. I was not entirely surprised. In fact, it replicated my experience from a few weeks earlier when, in a slightly impulsive experiment, I had used a set of ekphrases (more on this shortly) from the fifth century as prompts for Dall-E 2, only to find that many

of them likewise contained prohibited words. The ekphrases are a collection of forty-eight short descriptions of paintings from the Old and New Testaments that possibly adorned walls in late antiquity from a book called the *Dittochaeon*, written by the poet Prudentius around the year 400 CE. Many of the scenes Prudentius describes are well-known ones, such as the Nativity, the Adoration of the Magi, the Crucifixion, and the Resurrection. Prudentius's texts are formal, figurative, gestural, and antiquated. They are visually evocative, yet admittedly difficult to work with from the perspective of generative AI. Nonetheless, being a medievalist and an art historian, I was curious to see what a machine could make of these short, descriptive texts. I encountered two general results, both of which I found fascinating and troubling.

The first result was that the generators picked up on one or two key terms in the text, usually those at the very beginning, and resulted in a confused jumble. I did not demand a specific style; neither did I create iterations or variations. I simply wanted to see the first pass. For example, I entered Prudentius's description of Christ's baptism in the river Jordan:

> The Baptist, who fed on locusts and on honey from the woods and clothed himself in camel's hair, bathes his followers in the stream. He baptised Christ too, when suddenly the Spirit sent from heaven bears witness that it is He who forgives sin to the baptised who has himself been baptised.[2]

The resulting images were odd hodgepodges that picked up the terms "camel" and "baptism" and produced some amusingly odd creations (figs. 1 and 2).

A Della Francesca or Verrocchio these images were not, but I cannot fault the model too greatly for this. The above text is complex, and I used it as a challenge to what it could create when faced with something like this. One thing that struck me quite quickly, though, was how it conflated "camel's hair" with an actual camel—a piece of cloth for the animal itself.

The second general result was that Dall-E 2, like NightCafe a few weeks later, refused to create an image at all. When I entered Prudentius's description of the Adoration of the Magi—a scene found in Christmas displays across much of the world every year—I met a nearly blank screen with the small but stern words: "It looks like this request may not follow our content policy." What I had prompted was:

> Here the wise men bring costly gifts to the child Christ on the virgin's breast, of myrrh and incense and gold. The mother marvels at all the honours paid to the fruit of her pure womb, and that she has given birth to one who is both God and man and king Supreme.[3]

Again, this was a complex text for a prompt. And most people, if I might assume, would understand that the description is one of benign adoration: of a mother's love, a people's wonder, and the light of hope. Why then did Dall-E 2 refuse to create the image? Because it could not interpret Prudentius's figurative language. Specifically, the words "virgin," "breast," and "womb" ran afoul of OpenAI's safety policy. But in the context of the ekphrasis, Prudentius's text is not sexual in the least. Using "the virgin" as shorthand for the Virgin Mary is common enough; "breast" is used in the meaning of a mother holding a child against her breast or chest, as one does; and "womb" is not a gesture towards physiology, but rather highlights Mary's pure nature—indeed, "pure" is the most important modifier here. To see how deep the prohibition went, I modified the offending words one by one. First, I changed "virgin" to "the virgin Mary," but was still denied. I then changed "breast" to "lap," but could not get around "womb" being a prohibited term. For Dall-E 2 to make an image, I had to change "the virgin" to "Mary," "breast" to "lap," and remove "womb" altogether. The resulting images were vague and often ugly concoctions of nativity and adoration scenes (fig. 3).

The visual approximation was there, but, in order to get anywhere close to a legible version, I had to strip down Prudentius's language to conform to the prohibitions of OpenAI. In the process, the subtleties of evocation made way for blunt force description.

What interested me after all of this was 1) how text-to-image generators such as Dall-E 2, Midjourney, and others restrict the language of description and 2) how these restrictions determine the boundaries of generative image making, its ekphrastic nature, and the ensuing limits of AI's already mythologized emboldening of creativity. There are currently important discussions happening around many of the underpinnings of AI imagery, from unfair labor exploitation to copyright, intellectual property, plagiarism, embedded biases, and the opacity of how the models are trained. But here, I'd like to think a bit more about something that is very important to generative AI: the limitations of how it uses descriptive language in a literalist way as well as its ekphrastic foundations.

As I alluded to earlier, many commercial generative image models seem to pick up on a limited set of words when a complex description is entered as a prompt. This is perhaps not surprising. But what results is a theoretically important distinction between the short-form prompts that the models respond to and the subsequent way descriptive language is rendered in the model with actual descriptive language in this world. This distinction affects the way an image is generated—or, if an image can be generated at all, even when they point to the exact same thing. For example, when I entered "the crucifixion" as a prompt in Dall-E 2 and NightCafe, neither had problems generating pretty basic images. When

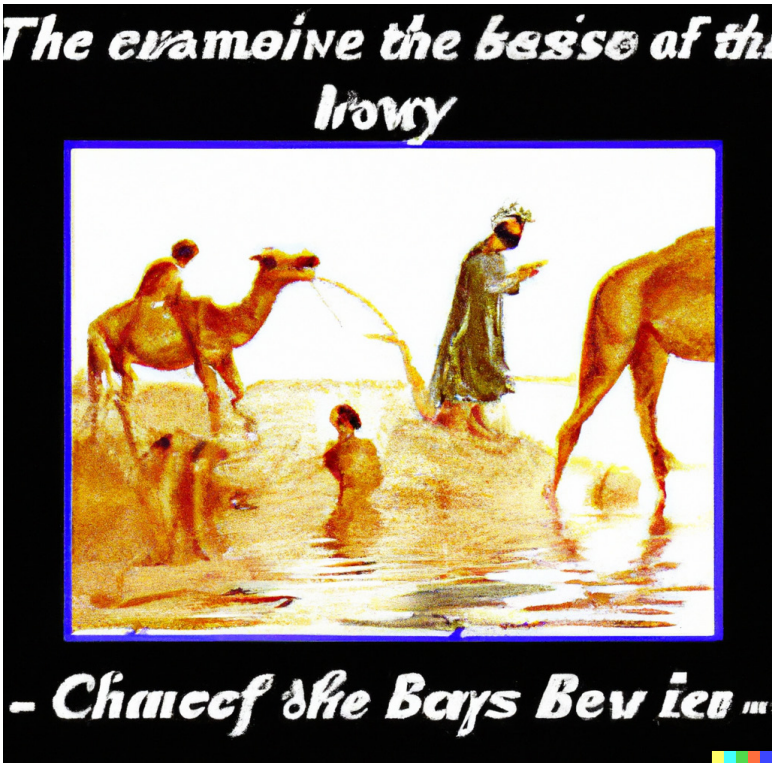

Figure 1: Image created in Dall-E 2 using the following text from Prudentius as the prompt: “The Baptist, who fed on locusts and on honey from the woods and clothed himself in camel’s hair, bathes his followers in the stream. He baptised Christ too, when suddenly the Spirit sent from heaven bears witness that it is He who forgives sin to the baptised who has himself been baptised.”

Figure 2: Image created in Dall-E 2 using the following text from Prudentius as the prompt: “The Baptist, who fed on locusts and on honey from the woods and clothed himself in camel’s hair, bathes his followers in the stream. He baptised Christ too, when suddenly the Spirit sent from heaven bears witness that it is He who forgives sin to the baptised who has himself been baptised.”

Figure 3: Image created in Dall-E 2 using the following text from Prudentius as the prompt (the offending words crossed out and/or replaced): "Here the wise men bring costly gifts to the child Christ on the virgin's [Mary's] breast [lap], of myrrh and incense and gold. The mother marvels at all the honours paid to the fruit of her pure womb, and that she has given birth to one who is both God and man and king Supreme."

Figure 4: Image created using Dall-E 2 with the prompt "man nailed to the cross."

I described the scene, however, things became unstable. For example, when I entered "man nailed to the cross," Dall-E 2 returned images of stock-photo-like men holding a cross with nails on it or, in one particularly interesting case, a man holding a cross made of nails (fig. 4).

I wondered if these confusions were the result of the model reading everything as a noun: man, nail (not the past-tense verb), and cross. A further attempt at detailed description caused Dall-E 2 to admonish me (fig. 5). What was the prompt? "A man nailed to a cross. Nails pierce his wrists. A wound is visible in his side." Simply, a description of the crucifixion.

The main point here is this: an image can be generated based on a prompt where terms associated with an image type can deliver the basic approximations, such as the crucifixion. But when the actual subject is described—when the objects that make up the generalized prompt are uttered—things fall apart. One enters a fuzzy quantum realm of descriptive prohibition that is difficult to unravel. We must ask: where exactly is this prohibitive line? It seems to be partly in terms of service. Dall-E 2, Midjourney, and NightCafe all have restricted sets of terms for prompting. OpenAI says that content should be "G-rated" and that terms that create violent, sexual, shocking, or political imagery, or that would otherwise "cause harm," are not allowed. This seems a reasonable position for a company to take. But what "causes harm" is anodyne, vague to the point of meaninglessness and effectively both an arbitrary and unaccountable decision. Moreover, not all text-to-image generators are that restrictive. With the open-source Stable Diffusion, one can imagine violent, racist, and sexual images to one's blackened heart's desire. But the deeper question remains: why is a violent image like the crucifixion allowed in a restrictive model like Dall-E 2 when called by title, but not when described?

I imagine another part of the answer lies in the way images and descriptions are transformed into vectors in the model. And these vectors are based on an understanding of descriptive language as mostly literal. Consequently, the only way to prompt for images is through a shrunken—if not vulgar—reduction in the means to describe an image. Not only are the images themselves being reduced in their visual and historical complexity, but description itself, the very way we communicate the intricacies of imagery, becomes leaden. If my thinking is correct—and I am sure there will be those eager to tell me the many ways in which I am wrong in this regard—I wonder from where this literal-mindedness of description is coming? We know that OpenAI uses millions of text-image pairs in their Contrastive Language-Image Pretraining (CLIP) model. But how were these pairs described in the first place? And who (or what) made that choice? These are important questions to ask, because the choices made have consequences.

Ekphrasis

The whole point is to think a bit harder about the linguistic limitations that generative AI operates under, which brings to me the question of ekphrasis. So let us get this out of the way: generative AI is an ekphrastic endeavor. It does what

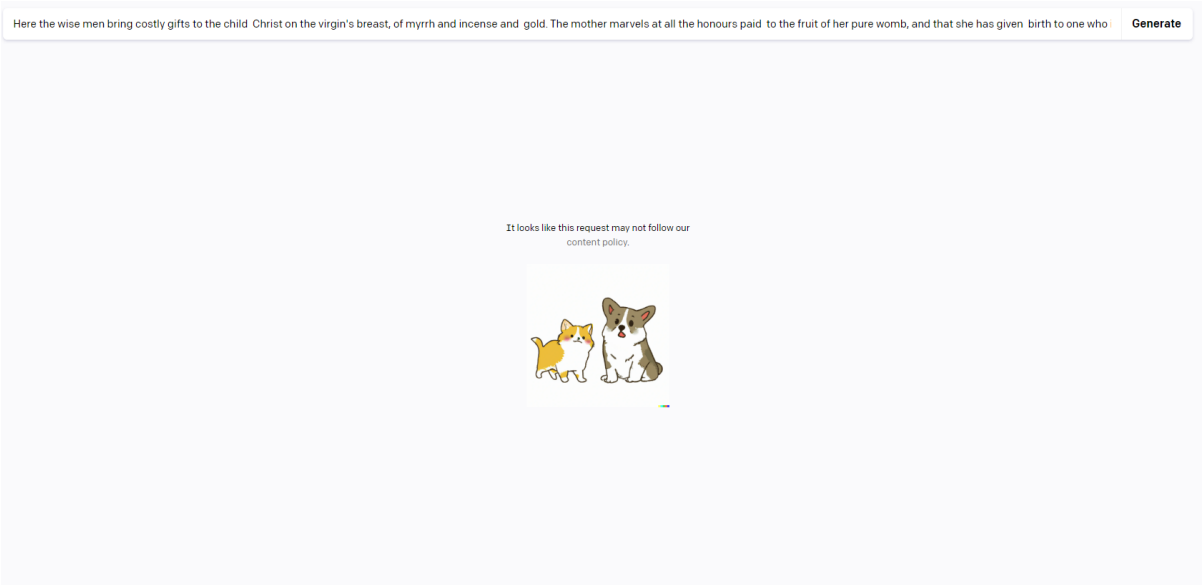

Figure 5: OpenAI's cutesy warning that a basic description of the crucifixion runs afoul of their content policy.

ekphrasis does at its most basic level: creates an image through verbal or textual description. Even the starting prompt for Midjourney nods towards ekphrasis: */imagine*. This is not so different from Prudentius or the second-century Greek sophist Philostratus (both the elder and the younger) imploring one to "see" an image through the ekphrastic art of description.[4] In fact, Philostratus's famous compilation of ekphrasis is titled *Imagines* (*Εikones* in Greek), sharing the same basic imperative as Midjourney's prompt. But here the machine stands at the center of ekphrasis, approximating a description into something more tangible. The machine fills in the gap based on just a few words. Prudentius and Philostratus had to explain what they wanted you to see and how they expected you to react. One had to let them be one's guide and meet them at their propositions.

However, while text-to-image generators are ekphrastic, they are only so in a limited and literal sense. In its current conception, ekphrasis often means describing an object—typically an artwork—that exists in the real world. But that is not ekphrasis in its more ancient or complex sense, where it refers to an advanced rhetorical tool used to convince someone of an object's existence in their mind's eye.[5] It is not only about describing form or content; ekphrasis attempts to communicate things like context, purpose, morals, surface, materials, meaning, interaction, reaction, quality, artistic skill, tone, and prosody. Ekphrasis is a rhetorical guide, but it is not told with the literal-mindedness of description that prompting demands.

We see that generative AI or text-to-images generators are ekphrastic *stricto sensu*. A person enters a short descriptive prompt for something that they imagine, and an operation ensues—the philosopher Hannes Bajohr reasonably calls this "operative ekphrasis."[6] The resulting images, as Roland Meyer points out, are recombinations of statistical precedence—imagery from the archive.[7] We end up with something of a median or composite-image type that reveals a common-denominator visual language, which is itself restricted by literal and categorical descriptive language. What this means is that the boundaries of descriptive language do not produce an unlimited imaginative range, but instead presents a restricted descriptive framework with its own assumptions about how images look and what they show. One must work within that. As such, it requires one to limit one's own imagination to the unarticulated assumptions of image classifications, and all that comes with it.

These assumptions are further obscured by some of the promises (whether marketing hype or not) of generative AI to create "realistic" or "accurate" images. What does it mean to create a "realistic" image? What are the required properties for an image to be realistic? To call an image "realistic" is not an obvious thing; to promise that generative imagery will be realistic is ultimately odd and unfalsifiable. The same questions arise for the promise of generative AI's ability to create an "accurate" image from a prompt. If the goal is to translate one's imagination via descriptive language as statistics, how can that be measured? And what if the output does not look the way one might imagine it? Is it accurate then? So goes the game and the limits of description.

There are some tools that allow us a better glimpse into the understanding of what these assumptions might be and what descriptive language follows. For example, with Hugging Face's CLIP-Interrogator tool, one can upload an image and in return get a prompt that would generate a similar image. When I uploaded a print by Albrecht Dürer depicting Christ

getting arrested, among the prompts that I received was: "a black and white drawing of a group of people, attack, albrecht durer, nazi propaganda," and so on. Never mind the fact that a print is not a drawing, why the tool churned out "nazi propaganda" as part of the prompt is disturbing and unclear. Is Dürer somehow, somewhere associated with Nazism in the model? If so—actual history aside—why? If not, why did this prompt associate this image with it? When I used the prompt to create an image, I received a jumble of black-and-white figures, vaguely Dürer-esque in the same way I would get from spilling coffee on a napkin and looking at the resulting shapes. Thankfully, no Nazis. Similarly, when I uploaded a Goya print of a bull fight, some of the prompts I received were "a black and white drawing of people and animals, including a matador & a bull, post game," and when I entered this prompt I received a stadium full of cows. Even if the results are peculiar, it still gives some insight into the way description- and image-generation are functioning.

The prompts themselves, however, show the limited imaginative range or expanse of text-to-image generators. And this imaginative restriction is furthered on two points. First, as we saw, is that there is a fault line between the short, literal language used for prompts with the more expressive, figurative and symbolic language of ekphrasis and description. To tap into generative imagery, one must stick to short descriptions and allow the machine to interpolate the rest. But second, what even makes a prompt or description detailed is not in itself clear. A recent paper on prompting estimates a detailed description is around eight words.[8] This is hardly descriptive in a traditional sense, but perhaps that is the point—and it makes sense given that text-to-image generators simply cannot, at present, handle overly detailed prompts. It nonetheless creates a very limited frame in which one can even imagine describing an image. As such, the *enargeia*—the imaginative vividness and presence that ekphrasis can unleash—remains locked-away behind verbal banality.

This, in my view, is perhaps much more confining than it is freeing as an artistic practice. Because one has to imagine how the machine might imagine something, one must limit one's words, in both kind and quantity, accordingly. Prompting thus becomes somewhat less reminiscent of ekphrasis and more of Wittgenstein's *Sprachspiele*: a language game in which one has to know the rules set by the algorithm.[9] To play the game—to get the image—one has to know that presently—although this is also changing—only short, literal declarations will do. And it might take many, many rounds of the game to "win"(in the sense that a generated image gets you what you want). Indeed, Bajohr has raised the point that images produced by generative AI can take hours and hours of iterative prompting to produce an image that satisfies one's own vision.[10] To enter and exit the game, one has to conform the limits of one's imagination to the limits of the model's approximations.

Creativity

Where, then, might the claim to creativity lay within the literalist approximations of generative imagery? I believe it is more in the mashing, mixing, and remixing of seemingly incongruous visual elements than anything else. And this can be quite fun—imagine a cat holding a sword in the style of Van Gogh. But I think this delight in incongruity might be mistaken for creativity. For this incongruity is based on the approximations of literal descriptive language and all its inherited limitations. Perhaps the output is already determined. The only question is what statistical recombination of the already-is will appear. But, one might argue, isn't that the case with any creative endeavor? Are we not so far off from a kind of techno-Dadaism of free play and playful unexpectedness? Are we not just taking what exists, vectorizing, and transforming it? And the answer is yes. . .to a point. One cannot escape history. But neither is one confined to its limitations in the way that one is with text-to-image generators. Art should not be determined by the determinations of description. Will, choice, play, and the vagueness and inscrutability of language all have their roles to play inside and outside the model.

Then again, at what point might this kind of work rise to the level of creativity rather than recombination? I do not have a good answer to this. This essay might be skeptical, but it is not dismissive. The question must therefore be posed nonetheless. And then, if we think of literal descriptive or categorized language as a—if not the—determining factor of generative imagery, what do we lose when we lack fundamental characteristics such as prosody? Tone? Sarcasm? Irony? How can a machine take into account all of the nonliteral mechanics of language that are so instrumental to both verbal and visual communication? Maybe someone somewhere else knows the answer. Maybe it simply does not matter whether the models will be able to encompass more nuance of descriptive language, whether the models become more adept at figurative language—this, no doubt, will change (as promised with ChatGPT's integration with Dall-E 3). Nevertheless, I hope my basic point remains valid regarding the use of these tools (or whatever they might be) and the limitations of the imagination and its imagining. Perhaps it is part of generative AI's nature not to even be concerned with such things. Yet that is the danger of mistaking a kind of magical-literalist approach to prompting and generating imagery for a genuinely creative process.

NOTES

1 Using the Stable algorithm and Stable Diffusion 1.5 model.

2 Prudentius, *Against Symmachus 2. Crowns of Martyrdom. Scenes From History. Epilogue*., trans. H. J. Thomson, Loeb Classical Library 398, (Cambridge: Harvard University Press, 1953). Available here: https://archive.org/details/imagines00philuoft/

3 Prudentius, *Against Symmachus 2. Crowns of Martyrdom. Scenes From History. Epilogue*., 359.

4 Prudentius, *Against Symmachus 2. Crowns of Martyrdom. Scenes From History. Epilogue*, 359.

5 For a historical definition of ekphrasis, see Ruth Webb, "Ekphrasis," Grove Art Online, https://doi.org/10.1093/gao/9781884446054.article.T025773.

6 Hannes Bajohr, "Operative Ekphrasis: The Collapse of the Text/Image Distinction in Multimodal AI," unpublished manuscript, last modified July 2023, PDF.

7 Roland Meyer, "The New Value of the Archive. AI Image Generation and the Visual Economy of 'Style,'" *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft* 37, no. 1 (2023): 100–11, http://dx.doi.org/10.25969/mediarep/22314.

8 Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen, "Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering," preprint version 2 (December 3, 2023), https://doi.org/10.48550/arXiv.2303.13534.

9 This idea is indebted to Marcus du Sautoy's thoughts on the limits of machine creativity. See Marcus du Sautoy, *The Creativity Code: Art and Innovation in the Age of AI* (Cambridge: Harvard University Press, 2019), 256.

10 See, for example, Hannes Bajohr, "Algorithmic Empathy: Toward a Critique of Aesthetic AI," *Configurations* 30, no. 2 (2022): 203–31.

BIBLIOGRAPHY

Bajohr, Hannes. "Algorithmic Empathy: Toward a Critique of Aesthetic AI." *Configurations* 30, no. 2 (2022): 203–31.

Bajohr, Hannes. "Operative Ekphrasis: The Collapse of the Text/Image Distinction in Multimodal AI." Unpublished manuscript, last modified July 2023. PDF.

Du Sautoy, Marcus. *The Creativity Code: Art and Innovation in the Age of AI.* Cambridge: Harvard University Press, 2019.

Meyer, Roland. "The New Value of the Archive. AI Image Generation and the Visual Economy of 'Style,'" *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft* 37, no. 1 (2023): 100–11. http://dx.doi.org/10.25969/mediarep/22314.

Oppenlaender, Jonas, Rhema Linder, and Johanna Silvennoinen. "Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering." Preprint version 2, submitted December 3, 2023. https://arxiv.org/abs/2303.13534.

Prudentius. Against Symmachus 2. Crowns of Martyrdom. Scenes From History. Epilogue. Translated by H. J. Thomson. Loeb Classical Library 398. Cambridge: Harvard University Press, 1953.

Webb, Ruth. "Ekphrasis." Grove Art Online. https://doi.org/10.1093/gao/9781884446054.article.T025773.

BENJAMIN ZWEIG is Project Manager for Digital Projects at Columbia University Libraries and Visiting Assistant Professor at the Pratt School of Information. He was previously the Digital Projects Coordinator at the National Gallery of Art and Research Associate for Digital Art History at the Center for Advanced Study in the Visual Arts. A medievalist by training, he received his Ph.D. in the history of art from Boston University, a M.A. in art history from Tufts University, and a B.F.A. in studio art from Massachusetts College of Art and Design. He has published widely on digital art history and medieval art, and has been the recipient of awards from he has been the recipient of awards from Boston University, the Kress Foundation, the Society for the Advancement of Scandinavian Study, and the Fulbright program.

Correspondence email: Benjamin.Zweig@gmail.com