media/rep/

Repositorium für die Medienwissenschaft

Eryk Salvaggio

How to Read an AI Image: Toward a Media Studies Methodology for the Analysis of Synthetic Images

2023

https://doi.org/10.25969/mediarep/22328

Veröffentlichungsversion / published version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Salvaggio, Eryk: How to Read an AI Image: Toward a Media Studies Methodology for the Analysis of Synthetic Images. In: *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft*. Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production, Jg. 19 (2023), Nr. 1, S. 83–99. DOI: https://doi.org/10.25969/mediarep/22328.

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under a Deposit License (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual, and limited right for using this document. This document is solely intended for your personal, non-commercial use. All copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute, or otherwise use the document in public.
By using this particular document, you accept the conditions of use stated above.

DFG Deutsche Forschungsgemeinschaft

Philipps Universität Marburg

IMAGE
The Interdisciplinary Journal of Image Sciences
37(1), 2023, S. 83-99
ISSN 1614-0885
DOI: 10.1453/1614-0885-1-2023-15456

HERBERT VON HALEM VERLAG

Eryk Salvaggio

How to Read an AI Image: Toward a Media Studies Methodology for the Analysis of Synthetic Images

Abstract: Image-generating approaches in machine learning, such as GANs and Diffusion, are actually not generative but predictive. AI images are data patterns inscribed into pictures, and they reveal aspects of these image-text datasets and the human decisions behind them. Examining AI-generated images as 'infographics' informs a methodology, as described in this paper, for the analysis of these images within a media studies framework of discourse analysis. This paper proposes a methodological framework for analyzing the content of these images, applying tools from media theory to machine learning. Using two case studies, the paper applies an analytical methodology to determine how information patterns manifest through visual representations. This methodology consists of generating a series of images of interest, following Roland Barthes' advice that "what is noted is by definition notable" (BARTHES 1977: 89). It then examines this sample of images as a non-linear sequence. The paper offers examples of certain patterns, gaps, absences, strengths, and weaknesses and what they might suggest about the underlying dataset. The methodology considers two frames of intervention for explaining these gaps and distortions: Either the model imposes a restriction (content policies), or else the training data has included or excluded certain images, through conscious or unconscious bias. The hypothesis is then extended to a more randomized sample of images. The method is illustrated by two examples. First, it is applied to images of faces produced by the StyleGAN2 model. Second, it is applied to images of humans kissing created with DALL·E 2. This allows us to compare GAN and Diffusion models, and to test whether the method might be generalizable. The paper draws some conclusions to the hypotheses generated by the method and presents a final comparison to an actual training dataset for StyleGAN2, finding that the hypotheses were accurate.

Background

Every AI-generated image is an *infographic* about the underlying dataset. AI images are data patterns inscribed into pictures, and they tell us stories about these image-text datasets and the human decisions behind them. As a result, AI images can become readable as 'texts'. The field of media studies has acknowledged "culture depends on its participants interpreting meaningfully what is around them [...] in broadly similar ways" (HALL 1997: 2). Images draw their power from intentional assemblages of choices, steered toward the purpose of communication. Roland Barthes suggests that images draw from and produce *myths*, a "collective representation" which turns "the social, the cultural, the ideological, and the historical into the natural" (BARTHES 1977: 165). Such myths are encoded into images by their creators and decoded by consumers (cf. HALL 1992: 117). For the most part, these assumptions have operated on the presumption that humans, not machines, were the ones encoding these meanings into images.

An AI has no unconscious mind, but nonetheless, contemporary Diffusion-based models produce images trained from collections of image-text pairings – datasets – which are produced and assembled by humans. The images in these datasets exemplify these collective myths and unstated assumptions. Rather than being encoded into the unconscious minds of the viewer or artist, they are inscribed into datasets. Machine learning models are meant to identify patterns in these datasets among vast numbers of images: DALL·E 2, for instance, was trained on 250 million text and image pairings (cf. RAMESH et al. 2021: 4). These datasets, like the images they contain, are created within specific cultural, political, social, and economic contexts. Machines are programmed in ways that inscribe and communicate the unconscious assumptions of human data-gatherers, who embed these assumptions into human-assembled datasets.

This paper proposes that when datasets are encoded into new sets of images, these generated images reveal layers of cultural and social encoding within the data used to produce them. This line of reasoning leads us to the research question: How might we read human myths through machine-generated images? In other words, what methods might we use to interrogate these images for cultural, social, political, or other artifacts? In the following, I will describe a loose methodology based on my training in media analysis at the London School of Economics, drawing from semiotic visual analysis. This approach is meant to "produce detailed accounts of the exact ways the meanings of an image are produced through that image" (ROSE 2012: 106). Rather than interpreting the images as one might an advertisement or film still, I suggest that AI images are best understood as infographics for their underlying dataset. The infographic, a fusion of *information* and *graphics*, has elsewhere been defined as the "visual representations of data, information, or concepts" (CHANDLER/MUNDAY 2011: 208)

that "consolidate and display information graphically in an organized way so a viewer can readily retrieve the information and make specific and/or overall observations from it" (HARRIS 1999: 198). The 'infographics' proposed here lack keys for interpreting the information they present because they are not designed to be interpreted as data but as imagery intended for human observers. Instead, we must use a semiotic analysis to reverse engineer the data-driven decisions that produced the image.

Conceptual Framework

The present paper proposes a methodology to understand, interpret, and critique the 'inhuman' outputs of generative imagery through a basic visual semiotic analysis as outlined in an introductory text by Gillian Rose (2001). It is intended to offer a similar introductory degree of simplicity. I began this work as an artist working with GANs in 2019, creating datasets – as well as images from these datasets. Through this work, I noticed patterns in the output, where information that was underrepresented in the dataset would be weakly defined in the corresponding images. Using StyleGAN to create diverse images of faces consistently produced more white faces than black ones. When black faces were generated, they lacked the definition of features found in white faces. This was particularly true for black women. In aiming to understand this phenomenon, I drew on media analysis techniques combined with an education in Applied Cybernetics, which examines complex systems through relationships and exchanges between components and their resulting feedback loops. While the present case studies examine the faces of black women in StyleGAN and images of men and women kissing in DALL·E 2, reflecting also on (the absence of) queer representations, the author is white and heterosexual. Any attempted determination of race, sexuality, or gender in AI-generated images inherently reflects this subjectivity.

Technical Background

Every image produced by diffusion models like DALL·E 2, Stable Diffusion, or Midjourney begins as a random image of Gaussian noise. When we prompt a Diffusion model to create an image, it takes this static and tries to reduce it. After a series of steps, it may arrive at a picture that matches the text description of one's prompt. The prompt is understood as a caption, and the algorithm works to 'find' the image in random noise based on this caption. Consider the way we look for constellations in the nighttime sky: If I tell you a constellation is up there, you mind find it – even if it isn't. Diffusion models are designed to

find constellations among ever-changing stars. Diffusion models are trained by watching images decay. Every image in the data has its information removed over a sequence of steps. This introduces noise, and the model is designed to trace the dispersal of this noise (or diffusion, hence the name) across the image. The noise follows a Gaussian distribution pattern, and as the images break down, noise clusters in areas where similar pixels are clustered. In human terms, this is like raindrops scattering an ink drawing across a page. Based on what remains of the image, the trajectory of droplets and motion of the ink, we may be able to infer where the droplet landed and what the image represented before the splash.

A Diffusion model is designed to sample the images, with their small differences in clusters of noise, and compare them. In doing this, the model makes a map of how the noise came in: learning how the ink smeared. It calculates the change between one image and the next, like a trail of breadcrumbs that lead back to the previous image. It will measure what changed between the clear image and the slightly noisier image. If we examine images in the process, we will see clusters of pixels around denser concentrations of the image. For example, flower petals, with their bright colors, stay visible after multiple generations of noise have been introduced. Gaussian noise follows a loose pattern, but one that tends to cluster around a central space. This digital residue of the image is enough to suggest a possible starting point for generating a similar image. From that remainder, it can find correlations in the pathways back to similar images. The machine is accounting for this distribution of noise and calculating a way to reverse it.

Once complete, information about the way this image breaks apart enters into a larger abstraction, which is categorized by association. This association is learned through the text-image pairings of CLIP (DALL·E 2) or LAION (Stable Diffusion, Midjourney, and others). The category flowers, for example, contains information about the breakdown of millions of images with the caption "flowers". As a result, the model can work its way backward from noise, and if given this prompt, "flowers", it can arrive at some generalized representation of a flower common to these patterns of clustering noise. That is to say: it can produce a perfect stereotype of a flower, a representation of any central tendencies found within the patterns of decay. When the model encounters a new, randomized frame of static, it applies those stereotypes in reverse, seeking these central tendencies anew, guided by the prompt. It will follow the path drawn from the digital residue of these flower images. Each image has broken down in its own way, but they share patterns of breakdown: clusters of noise around the densest concentrations of pixels, representing the *strongest signal* within the original images. In figure 1, we see an image of flowers compared to the 'residue' left behind as it is broken down.

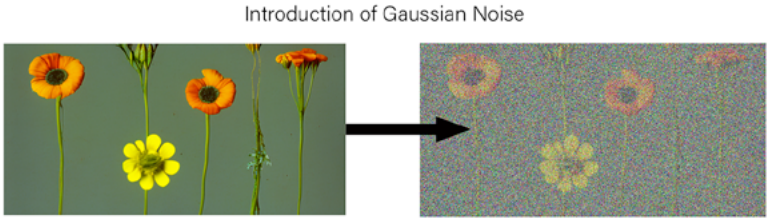


Figure 1: As Gaussian noise is introduced to the image, clusters remain around the densest concentrations of pixel information; created with Stable Diffusion in February 2023

As the model works backward from noise, our prompts constrain the possible pathways that the model is allowed to take. Prompted with "flowers", the model cannot use what it has learned about the breakdown of cat photographs. We might constrain it further: "Flowers in the nighttime sky". This introduces new sets of constraints: "Flowers", but also "night", and "sky". All of these words are the result of datasets of image-caption pairs taken from the world wide web. CLIP and LAION aggregate this information and then ignore the inputs. These images, labeled by internet users, are assembled into categories, or categories are inferred by the model based on its similarities to existing categories. All that remains is data – itself a biased and constrained representation of the social consensus, shaped by often arbitrary, often malicious, and almost always unconsidered boundaries about what defines these categories.

This paper proposes that when we look at AI images, specifically Diffusion images, we are looking at infographics about these datasets, including their categories, biases, and stereotypes. To read these images, we consider them representations of the underlying data, visualizing an 'internet consensus'. They produce images where prompts produce abstractions of centralizing tendencies. When images are more closely aligned to the abstract ideal of these stereotypes, they are clean, 'strong' images. When images drift from this centralizing consensus, they are more difficult to categorize. Therefore, images of certain categories may appear 'weak' – either occurring less often or with lower definition or clarity.

These ideal 'types' are socially constructed and encoded by anyone who uploads an image to the internet with a descriptive caption. For example, a random sample of the training data associated with the phrase "Typical American" within the LAION 5B dataset that drives Stable Diffusion suggests the images and associations for "Typical American" as a category: images of flags, painted faces from Independence Day events, as would be expected. Social stereotypes, related

to obesity and cowboy hats, are also prevalent. Curiously, one meme appears multiple times, a man holding a Big Gulp from 7-11 (a kind of large, frozen sugar drink). Figure 2 is an image in response to the prompt "Typical American" in which the man holds a large beverage container, like a Big Gulp, whilst wearing face paint and a cowboy hat. We see that while the relationship between the dataset and the images that Diffusion produces are not literal, these outcomes are nonetheless connected to the concepts tied to this phrase within the dataset.



Figure 2: A result from the prompt "Typical American" from Stable Diffusion in February 2023

Just as archives are the stories of those who curate them, Diffusion generated images are no different. They visualize the constraints of the prompt, as defined by a dataset of human-generated captions that is assembled by CLIP or LAION's automated categorizations. I propose that these images are a visualization of this archive. They struggle to show anything the archive does not contain or is not clearly categorized in accordance with prompts. This suggests that we can read images created by these systems. The next section proposes a methodology for reading these images which blends media analysis and data auditing techniques. As a case study, it presents DALL·E 2 generated images of people kissing.

Methodology

Here I will briefly outline the methodology, followed by an explanation of each step in greater detail.

1. Produce images until you find one image of particular interest.
2. Describe the image simply, making note of interesting and uninteresting features.
3. Create a new set of samples, drawing from the same prompt or dataset.

4. Conduct a content analysis of these sample images to identify strengths and weaknesses.
5. Connect these patterns to corresponding strengths and weaknesses in the underlying dataset.
6. Re-examine the original image of interest.

Each step is explained through a case study of an image produced through DALL·E 2. The prompt used to generate the image was "Photograph of two humans kissing". This prompt was used until an image of particular interest caught my eye. Each step is described, with further discussions of the step integrated into each section.



Figure 3: "Photograph of two humans kissing", produced with DALL·E 2 in February 2023

1. Produce Images until you Find one of Particular Interest

First, we require a research question. There is no methodology for selecting images of interest. Following Rose, images were chosen subjectively, "on the basis of how conceptually interesting they are" (ROSE 2012: 73). Images must be striking, but their relevance is best determined by the underlying question being pursued by the researcher. The case studies offered here were produced through simple curiosity. I aimed to see if a sophisticated AI models could create compelling images of human emotion. I began with the image displayed in figure 3.

2. Describe the Image Simply, Making Note of Interesting and Uninteresting Features

We need to know what is in the image in order to assess why they are there. In Case Study 1 (fig. 3), the image portrays a heterosexual white couple. A reluctant(?) male is being kissed by a woman. In this case, the man's lips are protruding, which is rare compared to our sample. The man is also *weakly* represented: his eyes and ears have notable distortions. In the following analysis of the image, *weak* features thus refer to smudged, blurry, distorted, glitched, or otherwise striking features of the image. *Strong* features represent aspects of the image that are of high clarity, realistic, or at least realistically represented.

While this paper examines photographs, similar weak and strong presence can be found in a variety of images produced through Diffusion systems in other styles as well. For example, if oil paintings frequently depict houses, trees, or a particular style of dress, it may be read as a strong feature that would be matched to a strong correspondence with aspects of the dataset. You may discover that producing oil paintings in the style of 18th century European masters does not generate images of black women. This would be a weak signal from the data, suggesting that the referenced datasets of 18th century portraiture did not contain portraits of black women (Note that these are hypotheticals and have not been specifically verified).

3. Create a New Set of Samples, Drawing from the Same Prompt or Database

Creating a wider variety of samples allows us to identify patterns that might reveal this central tendency in the abstraction of the image model. As the model works backwards from noise – following constraints on what it can find in that noise – we want to create many images to identify any gravitation toward its average representation. It is initially challenging to find insights into a dataset through a single image. However, generative images are a medium of scale: millions of images can be produced in a day, with streaks of variations and anomalies. None of these reflect a single author's choices. Instead, they blend thousands, even millions of aggregated choices. By examining the shared properties of many images produced by the same prompt or dataset, we can begin to understand the underlying properties of the data that formed them. In this sense, AI imagery may be analyzed as a series of film stills: a sequence of images, oriented toward 'telling the same story'. That story is the dataset. The dataset is revealed through a non-linear sequence, and a larger sample will consist of a series of images designed to tell that same story. Therefore, we would create variations

using the same prompt or model. I use a minimum of nine, because nine images can be placed side by side and compared on a grid. For some examinations, I have generated 18-27 or as many as 90-120. While creating this expanded sample set, we would continue to look for any *conceptually interesting* images from the same prompt. These images do not have to be notable in the same way that the initial source image was. The image that fascinated, intrigued, or irritated us was interesting for a reason. The priority is to understand that reason by understanding the *context* – interpreting the patterns present across many similarly generated images. We will not yet have a coherent theory of what makes these images notable. We are simply trying to understand the *generative space that surrounds the image of interest*. This generative, or latent space, is where the data's weaknesses and strengths present themselves. Even a few samples will produce recognizable patterns, after all.



Figure 4: Nine images created from the same prompt as our source image, created with DALL·E 2 in February 2023. If you want to generate your own, you can type "Photograph of humans kissing" into DALL·E 2 and grab samples for comparison yourself

4. Conduct a Content Analysis of these Sample Images to Identify Individual Strengths and Weaknesses

Now we can study the new set of images for patterns and similarities by applying a form of content analysis. We describe what the image portrays 'literally' (the denoted meaning). Are there particularly strong correlations between any of the images? Look for certain compositions/arrangements, color schemes, lighting effects, figures or poses, or other expressive elements, that are strong across all (or some meaningful subsections) of the sample pool. These indicate certain biases in the source data. When patterns are present, we will call these *signals*. Akin to symptoms, indicators are observable elements of the image that point to a common underlining cause. We may have strong signals: suggesting frequency

of the pattern in the data pattern, the strongest signals being near-universal and the strongest dismissed as obvious. A strong signal would include tennis balls being round, cats having fur, etc. A weak signal, on the other hand, suggests that the image is on the peripheral of the model's central tendencies for the prompt. The most obvious indicators of weak signals are images that simply cannot be created realistically or with great detail. The smaller the number of examples in a dataset, the fewer images the model may learn from, and the more errors will be present in whatever it generates. These may be visible in blurred appearances, such as smudges, glitches, or distortions. Weak signals may also be indicated through a comparison of what patterns are present against what patterns might otherwise be possible.

Strong signals: In the given example, the images render skin textures quite well. They seem professionally lit, with studio backgrounds. They are all close-ups focused on the couple. Women tend to have protruding lips, while men tend to have their mouths closed. These therefore suggest *strong* signals in the data, suggesting an adjacency to central tendencies within the assigned category of the prompt. These signals may not be consistent across all images, but are important to recognize because they provide a contrast and context for what is weakly represented.

Weak signals: In the case study, two important things are apparent to me. First, most pictures are heteronormative, i.e., the images portray only man/woman couples. The present test run, created in November 2022, differs from an earlier test set (created in October 2022 and made public online, cf. SALVAGGIO 2022). In the original test set, *all* couples were heterosexual. Second, there is a strong presence of multiracial couples: another change from October 2022 when nearly all couples shared skin tones. Third, they are missing convincing interpersonal contact. This is, in fact, identical in both datasets from different months. The *strong* signal across the kissing images might be a sense of hesitancy as if an invisible barrier exists between the two partners in the image. The lips of the figures are weak: inconsistent and imperfect. With an inventory of strong and weak patterns, we can begin asking critical questions toward a hypothesis.

1. What data would need to be present to explain these strong signals?
2. What data would need to be absent to explain these weak signals?

Weaknesses in your images may be a result of sparse training data, training biased toward exclusion, or reductive system interventions such as censorship. Strengths may be the result of prevalence in your training data, or encouraged by system interventions. They may also represent cohesion between your prompt and the 'central tendency' of images in the dataset, for example, if you prompt "apple", you may produce more consistent and realistic representations of apples than if you request an "apple-car". For example, DALL·E 2 introduces diversifying keywords randomly into prompts (cf. OFFERT/PHAN 2022). The more often some

feature is in the data, the more often it will be emphasized in the image. In summary, you can only see what's *in* the data and you cannot see what is *not* in the data. When something is strikingly wrong or unconvincing, or repeatedly impossible to generate at all, that is an insight into the underlying model.

An additional case study could provide even more context. In 2019, while studying the FFHQ dataset that was used to generate images of human faces for StyleGAN, I noted that the faces of black women were consistently more distorted than the faces of other races and genders. I asked the same question: What data was present to make white faces so clear and photorealistic? What data was absent to make black women's faces so distorted and uncanny? I began to formulate a hypothesis. In the case of black women's faces being distorted, I could hypothesize that black women were *underrepresented* in the dataset: that this distortion was the result of a weak signal. In the case study of kissing couples, something else is missing. One hypothesis might be that the dataset used by OpenAI does not contain many images of *anyone* kissing. That might explain the awkwardness of the poses. I might also begin to inquire about the absence of same-sex couples and conclude that LGBTQ couples were absent from the dataset. While unlikely, we may use this as an example of how to test that theory, or whatever you find in your own samples, in the next step.

5. Connect these Patterns to Corresponding Strengths and Weaknesses in the Underlying Dataset

Each image is the product of a dataset. To continue our research into interpreting these images, it is helpful to address the following questions as specifically as possible:

1. What is the dataset and where did it come from?
2. What can we verify what is included in the dataset and what is excluded?
3. How was the dataset collected?

Often, the source of training data is identified in white papers associated with any given model. There are tools being developed – such as Matt Dryhurst and Holly Herndon's *Swarm*, that can find source images in some sets of training data (LAION) associated with a given prompt. When training data is available, it can confirm that we are interpreting the image-data relationship correctly. OpenAI trained DALL·E 2 on hundreds of millions of images with associated captions. As of this writing, the data used in DALL·E 2 is proprietary, and outsiders do not have access to those images. In other cases, the underlying training dataset is open source, and a researcher can see what training material they draw from. For the sake of this exercise, we'll look through the LAION dataset, which is used for the diffusion engines Stable Diffusion and Midjourney. When we look at the images

that LAION uses for "Photograph of humans kissing", we can see that the training data for this prompt in that library consists mostly of stock photographs where actors are posed for a kiss, suggesting a database trained on images displaying a lack of genuine emotion or any romantic connection. For GAN models, which produce variations on specific categories of images (for example, faces, cats, or cars), many rely on open training datasets containing merely thousands of images. Researchers may download portions of them and examine a proportionate sample. This may become exponentially harder as datasets become exponentially larger. For examining race and face quality through StyleGAN, I downloaded the training data – the FFHQ dataset – and randomly examined a sub-portion of training images to look for racialized patterns. This confirmed that the proportion of white faces far outweighed faces of color.

While we do not have training data for DALL·E 2, we can make certain inferences by examining other large datasets. For example, we might test the likelihood of a hypothesis that the dominance of heterosexual couples in stock photography contributes to the relative absence of LGBTQ subjects in the images. This would explain the presence of heterosexual couples (a *strong signal* from the dataset) and the absence of LGBTQ couples that occurred in our earlier tests from 2022. However, LAION's images found for the prompt query "kissing" is almost exclusively pictures of women kissing. While DALL·E 2's training data remains in a black box, we now have at least some sense of what a large training set *might* look like and can recalibrate the hypothesis. The massive presence of women kissing women in the dataset suggests that the weak pattern is probably not a result of sparse training data or a bias in data. We would instead conclude that the bias runs the other way: if the training data is overwhelmed with images of women kissing, then the outcomes of the prompt should also be biased toward women kissing. Even in the October 2022 sample, however, women kissing women seemed to be rare in the generated output.

This suggests we need to look for *interventions*. An intervention is a system-level design choice, such as a content filter, which prevents the generation of certain images. Here we *do* have data even for DALL·E 2 that can inform this conclusion. 'Pornographic' images were explicitly removed from OpenAI's dataset to ensure it does not reproduce similar content. Other models, such as LAION, contain vast amounts of explicit and violent material (cf. BIRHANE 2021). By contrast, OpenAI deployed a system-level intervention into their dataset:

> We conducted an internal audit of our filtering of sexual content to see if it concentrated or exacerbated any particular biases in the training data. We found that our initial approach to filtering of sexual content reduced the quantity of generated images of women in general, and we made adjustments to our filtering approach as a result (OPENAI 2022: n.pag.).

Requests to DALL·E 2 are hence restricted to what OpenAI calls 'G-rated' content, referring to the motion picture rating for determining age appropriateness.

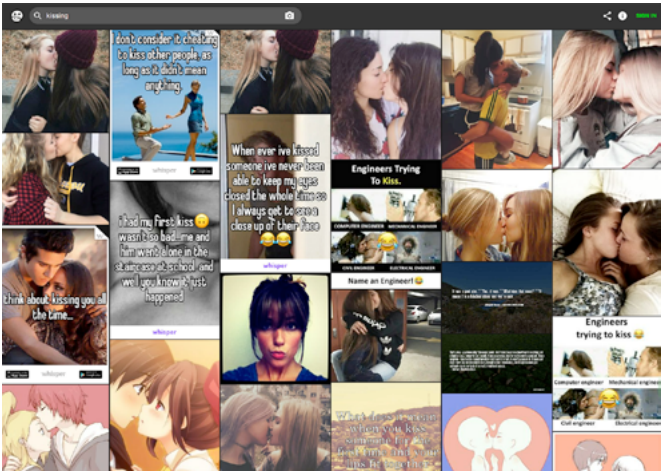


Figure 5: First page of screen results from a search of LAION training data associated with the word "Kissing" indicates a strong bias toward images of women kissing, Screen grab from haveibeentrained.com [Accessed March 22, 2023]

G-rated means appropriate for all audiences. The intervention of removing images of women kissing (or excluding them from the data-gathering process) as 'pornographic' content reduced references to women in the training data. The G-rating intervention could also explain the barrier effect between kissing faces in our sample images, a result of removing images where kissing might be deemed sexually charged. We may now begin to raise questions about the criteria that OpenAI drew around the notion of 'explicit' and 'sexual' content. This leads us to new sets of questions helpful to forming a consecutive hypothesis.

1. What are the boundaries between forbidden and permitted content in the model's output?
2. What interventions, limitations, and affordances exist between the user and the output of the underlying dataset?
3. What cultural values are reflected in those boundaries?

Next is to test these questions. One method is to test the limits of OpenAI's restricted content filter which prevents the completion of requests for images that depict pornographic, violent, or hateful imagery. Testing this content filter, it is easy to find out that a request for an image of "two men kissing" creates an image of two men kissing. Requesting an image of "two women kissing" triggers a warning for "explicit" content (this is true as of February 2023). This offers a clear example of mechanisms through which cultural values become inscribed into AI image production. First, through the dataset: what is collected, retained,

and later trained on. Second, through system-level affordances and/or interventions: what can and cannot be produced or requested.

6. Re-examine the Original Image of Interest

We now have a hypothesis for understanding our original image. We may decide that the content filter excludes women kissing women from the training data as a form of 'explicit' content. We deduce this because women kissing is flagged as explicit content on the output side, suggesting an ideological, cultural, or social bias against gay women. This bias is evidenced in at least one content moderation decision (banning their generation) and may be present in decisions about what is and is not included in the training data. The strangeness of the pose in the initial image, and of others showing couples kissing, may also be a result of content restrictions in the training data that reflect OpenAI's bias toward, and selection for, G-rated content. How was 'G-rated' defined, however, and how was the data parsed from one category to another? Human, not machinic, editorial processes were likely involved. Including more 'explicit' images in the training model likely wouldn't solve this problem – or create new ones. Pornographic content would create additional distortions. But in a move to exclude explicit content, the system has also filtered out women kissing women, resulting in a series of images that recreate dominant social expectations of relationships and kisses as 'normal' between men and women.

Returning to the target image, we may ask: What do we see in it that makes sense compared to what we have learned or inferred? What was encoded into the image through data and decisions? How can we make sense of the information encoded into this image by the data that produced it? With a few theories in mind, I would run the experiment again: this time, rather than selecting images for the patterns they shared with the notable image, use any images generated from the prompt. Are the same patterns replicated across these images? How many of these images support the theory? How many images challenge or complicate the theory? Looking at the broader range of generated images, we can see if our observations apply consistently – or consistently enough – to make a confident assertion. Crucially, the presence of 'successful' images does not undermine the claim that weak images reveal weaknesses in data. Every image is a statistical product: odds are weighted toward certain outcomes. When you see successful outcomes fail, that failure offers *insight* into gaps, strengths, and weaknesses of those weights. They may occasionally – or predominantly – be rendered well. What matters to us is what the failures suggest about the underlying data. Likewise, conducting new searches across time can be a useful means of tracking evolutions, acknowledgments, and calibrations for recognized biases.

As stated earlier, my sampling of AI images from DALL·E 2 conducted showed swings in bias from predominantly white, heterosexually coded images toward greater representations of genders and skin tones.

Finally, we may conclude that AI generated images of couples kissing is the result of technical limits. Lips kissing may reflect a well-known flaw in rendering human anatomy. Both GANs and Diffusion models, for example, frequently produce hands with an inappropriate number of fingers. There is no way to constrain the properties of fingers, so they can become tree roots, branching in multiple directions, multiple fingers per hand with no set length. Lips, too, can seem to be more constrained, but the variety and complexity of lips, especially in contact with each other, may be enough to distort the output of kissing prompts. Hands and points of contact between bodies – especially where skin is pressed or folds – are difficult to render well.

Discussion & Conclusion

Each of these hypotheses warrants a deeper analysis than the scope of this paper would allow. The goal of this paper was to present a methodology toward the analysis of generative images produced by Diffusion-based models. Our case study suggests that examples of cultural, social, and economic values are embedded into the dataset. This approach, combined with more established forms of critical image analysis, can give us ways to read the images as infographics. The method is meant to generate insights and questions for further inquiry, rather than producing statistical claims, though one could design research for quantifying the resulting claims or hypotheses. The model has succeeded in generating strong claims for further investigations interrogating the underlying weaknesses of image generation models. This includes the absence of black women in training datasets for StyleGAN, and now, the exclusion of gay women in DALL·E 2's output. Ideally, these insights and techniques move us away from the 'magic spell' of spectacle that these images are so often granted. It is intended to provide a deeper literacy into where these images are drawn from. Identifying the widespread use of stock photography, and what that means about the system's limited understanding of human relationships, emotional and physical connections, is another pathway for critical analysis and interpretations.

The method is meant to move us further from the illusion of 'neutral' and unbiased technologies which is still prevalent in the discourse around these tools. We often see AI systems deployed as if they are free of human biases – the Edmonton police (Canada) recently issued a wanted poster including an AI-generated image of suspect based on his DNA (cf. XIANG 2022). That's pure mystification. They are *bias engines*. Every image should be read as a map of those biases,

and they are made more legible using this approach. For artists and the general public creating AI-images, it also points to a strategy for revealing these problems. One constraint of this approach is that models can change at any given time. It is obvious that OpenAI could recalibrate their DALL·E 2 model to include images of women kissing tomorrow. However, when models calibrate for bias on the user end it does not erase the presence of that bias. Models form abstractions of categories based on the corpus of the images they analyze. Removing access to those images, on the users end, does not remove their contribution to that abstraction. The results of early, uncalibrated outcomes are still useful in analyzing contemporary and future outputs. Generating samples over time also presents opportunities for another methodology, tracking the evolution (or lack thereof) for a system's stereotypes in response to social changes. Media studies may benefit from the study of models that adapt or continuously update their underlying training images or that adjust their system interventions.

Likewise, this approach has limits. One critique is that researchers cannot simply look at training data that is not accessible. As these models move away from research contexts and toward technology companies seeking to make a profit from them, proprietary models are likely to be more protected, akin to trade secrets. We are left making informed inferences about DALL·E 2's proprietary dataset by referencing datasets of a comparable size and time frame, such as LAION 5B. Even when we can find the underlying data, researchers may use this method only as a starting point for analysis. It raises the question of where to begin even when there are billions of images in a dataset. The method marks only a starting point for examining the underlying training structures at the site where audiences encounter the products of that dataset, which is the AI-produced image.

Thanks to Valentine Kozin and Lukas R.A. Wilde for feedback on an early draft of this essay.

Bibliography

BARTHES, ROLAND: *Image, Music, Text.* Translated by Stephen Heath. London [Fontana Press] 1977

BIRHANE, ABEBA; VINAY UDAY PRABHU; EMMANUAEL KAHEMBWE: Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes. *arXiv:2110.01963*. October 5, 2021. https://arxiv.org/abs/2110.01963 [accessed February 16, 2023]

CHANDLER, DANIEL; ROD MUNDAY: *A Dictionary of Media and Communication*. Oxford [Oxford University Press] 2011

HALL, STUART: Encoding/Decoding. In: *Culture, Media, Language: Working Papers in Cultural Studies, 1972-1979*. London [Routledge] 1992, pp. 117-127

HALL, STUART: The Work of Representation. In: *Representation: Cultural Representations and Signifying Practices*. London [Sage] 1997, pp. 15-74

HARRIS, ROBERT: *Information Graphics: A Comprehensive Illustrated Reference*. New York [Oxford University Press] 1999

OPENAI: DALL·E 2 Preview – Risks and Limitations. In: *GitHub*. July 19, 2022. https://github.com/openai/dalle-2-preview/blob/main/system-card.md [accessed February 16, 2023]

OFFERT, FABIAN; THAO PHAN: A Sign That Spells: DALL-E 2, Invisual Images and the Racial Politics of Feature Space. *arXiv:2211.06323*. October 26, 2022. https://arxiv.org/abs/2211.06323 [accessed February 20, 2023]

RAMESH, ADITYA; et al.: Zero-Shot Text-to-Image Generation. *arXiv:2102.12092*. February 24, 2021. https://arxiv.org/abs/2102.12092 [accessed February 16, 2023]

ROSE, GILLIAN: *Visual Methodologies: An Introduction to Researching with Visual Materials*. London [Sage] 2001

SALVAGGIO, ERYK: How to Read an AI Image: The Datafication of a Kiss. In: *Cybernetic Forests*. October 2, 2022. https://cyberneticforests.substack.com/p/how-to-read-an-AI-image [accessed February 16, 2023]

XIANG, CHLOE: Police are Using DNA to Generate Suspects they've Never Seen. In: *Vice Media*. October 11, 2022. https://www.vice.com/en/article/pkgma8/police-are-using-dna-to-generate-3d-images-of-suspects-theyve-never-seen [Accessed February 18, 2023]