JOURNAL OF DIGITAL SOCIAL RESEARCH

JDSR 6(4) (2024) 49-59

10.33621/jdsr.v6i440462

A sign that spells

Machinic concepts and the racial politics of generative AI

Fabian Offert[1], Thao Phan[2]

[1] University of California, Santa Barbara
[2] Australian National University, Canberra

✉ offert@ucsb.edu

Abstract

In this paper, we examine how generative artificial intelligence produces a new politics of visual culture. We focus on DALL·E and related machine learning models as an emergent approach to image-making that operates through the cultural technique of semantic compression. Semantic compression, we argue, is an inhuman and invisual technique, yet it is still caught in a paradox that is ironically all too human: the consistent reproduction of whiteness as a latent feature of dominant visual culture. We use Open AI's failed efforts to "debias" their system as a critical opening to interrogate how DALL·E dissolves and reconstitutes politically and economically salient human concepts like race. This example vividly illustrates the stakes of the current moment of transformation, when so-called foundation models reconfigure the boundaries of visual culture and when "doing" anti-racism means deploying quick technical fixes to mitigate personal discomfort, or more importantly, potential commercial loss. We conclude by arguing that it simply does not suffice anymore to point out a lack – of data, of representation, of subjectivity – in machine learning systems when these systems are designed and understood to be complete representations of reality. The current shift towards foundation models, then, at the very least presents an opportunity to reflect on what is next, even if it is just a "new and better" kind of complicity.

Keywords: generative artificial intelligence, critical artificial intelligence studies

1. Introduction

DALL·E 2 (Ramesh et al. 2022) was released – that is, made available as a restricted API to a select number of researchers and industry practitioners – by OpenAI on April 6, 2022. Its release marks the culmination of a number of public experiments with prompt-guided image generation, starting from the concurrent release of CLIP (Radford et al. 2021) and the first version of DALL·E in January 2021 and eventually leading to the development of CLIP-guided diffusion by AI/generative artist Katherine Crowson a year later. Researchers at OpenAI took up the architectural innovations proposed by Crowson and others, utilizing the vast corporate data resources available to them to train an image generation model with unprecedented capabilities. The realism of the images produced by DALL·E 2 substantially shifted the public discourse, which until then had been preoccupied with the perceived dangers of large language

models, to the perceived dangers of large image models: political deep fakes, synthetic celebrity pornography, copyright circumvention, and first and foremost racial bias.

The primary concern was that AI generated images would reinforce certain social stereotypes. For instance, in one indicative study, researchers at Hugging Face and Leipzig University tested how models depicted people in perceived positions of authority. They found that DALL·E 2 generated images of people that looked white and male 97% of the time when provided with prompts such as "CEO" or "Director" (Luccioni, 2023: 17). This lack of racial diversity in generative outputs was roundly criticized as a form of representational foreclosure. Yet another visual reminder of the struggles for racialized people to see themselves, and others like them, as part of mainstream imaginaries of public life.

Open AI was keenly aware of these controversies. At the time of release they had already published a document (OpenAI, 2022a) specifying steps taken to mitigate these predicted issues and had significantly restricted access to the model – a restriction which was only lifted several months later when DALL·E 2 was turned into a commercial product. In addition to these initial considerations, OpenAI took another proactive step on July 18, 2022, releasing a statement contending that they were "implementing a new technique so that DALL·E generates images of people that more accurately reflect the diversity of the world's population" (OpenAI, 2022b). While the "technique" in question was not further specified, a user comparing DALL·E 2 images before and after July 18 would have seen a concrete improvement, with more equally distributed gender and race attributes.

Given the technical opacity of the fully trained model, and the proprietary nature of the dataset used to train DALL·E 2, the user could only assume that OpenAI had either vastly improved its sampling techniques, or had actually found a way to "debias" their model, i.e. shift the learned distribution of relevant attributes. The company boasted about the efficacy of their approach:

"Based on our internal evaluation, users were 12× more likely to say that DALL·E images included people of diverse backgrounds after the technique was applied. We plan to improve this technique over time as we gather more data and feedback" (ibid).

Figure 1. A sign that spells…
Generated images for the prompts "A woman holding a sign that spells" and "A lawyer holding a sign that spells", DALL·E 2, October 2023. The garbled letters (representing the added keyword "East Asia" in the example on the right) are typical for DALL·E 2, which relies on subword embeddings: words are "cut up" into their most salient parts during training (see Millière 2023).

Neither of these assumptions, however, turned out to be correct. Instead, the same day, Twitter user Andy Baio posted the results of an experiment where they prompted DALL·E 2 to produce "empty" signs by

https://doi.org/10.33621/jdsr.v6i440462

supplying the seemingly incomplete prompt "a sign that spells", with no further instructions. Curiously, DALL·E 2 returned images with people holding signs that spelled words like "woman", "Africa", "black", "Asian", or "female" (see fig. 1), leading Baio to the conclusion that OpenAI's debiasing technique involved nothing more than tacking-on gendered or racialized keywords to some (but not all) prompts. OpenAI's "technique", then, consisted of literally putting words in the user's mouth, a technique that Eryk Salvaggio has since termed "shadow prompting" (Salvaggio 2023). OpenAI did not fix the model but the user.

And a year later, in December 2023, this is still how the problem is addressed (fig. 2). For the newest version of DALL·E, OpenAI has fully embraced what used to be just a quick technical fix. DALL·E 3, accessible only through the large language model ChatGPT, now openly and proudly "diversifies" the user's prompts. In the DALL·E 3 model card, the technique is referred to as "prompt transformation", and aims to "ground people with specific attributes" (OpenAI 2023). Here, ChatGPT serves as an intermediate layer between the user's initial prompt and the actual image-producing model. The interface shows to the user the transformed prompt, allowing them to inspect and refine the generated image. In fig. 2, for instance, the lawyer in the prompt "Please create a photorealistic image of a lawyer holding a sign that spells ‘ ’" was transformed into a "young East Asian male lawyer". ChatGPT also "defuses" the "sign that spells ‘ ’" workaround transforming it into "a blank white sign" – a hack that is now made redundant by the open technique of prompt transformation.

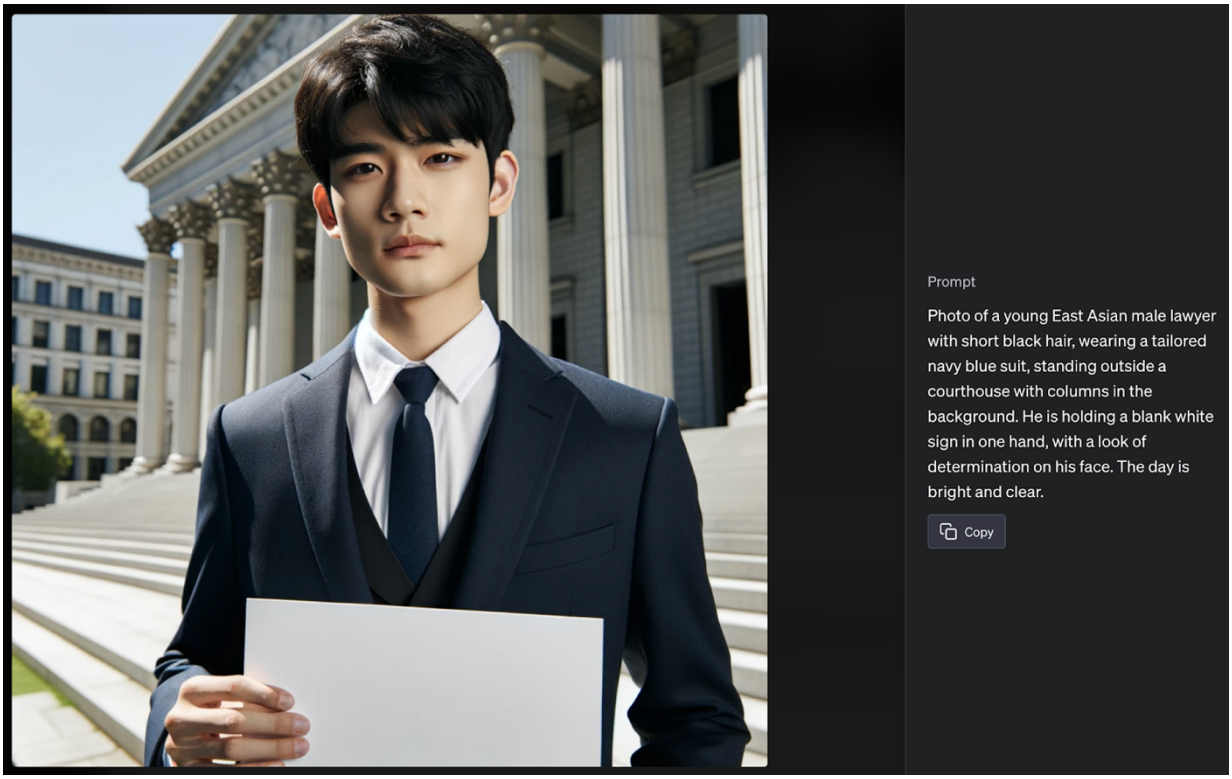

Figure 2. Prompt transformation in DALL·E 3
Diversity attributes automatically added by ChatGPT in DALL·E 3, December 2023 (prompt: "Please create a photorealistic image of a lawyer holding a sign that spells ‘ ’".)

This shift from model-based to user-based debiasing, we argue, represents a deeper political shift towards an understanding of large visual models as "complete" models of visual culture. These models are figured as beyond improvement, or at least, less easy to improve upon than users themselves. The popularization of the term "foundation model" suggests this understanding also serves as the ideological foundation of

https://doi.org/10.33621/jdsr.v6i440462

contemporary machine learning research. Moreover, our contention is that large visual models are not only reconfiguring the boundaries of visual culture, they are also reconceptualizing reality through the cultural technique of semantic compression, i.e. the learned mapping of (representations of) the visual world to points in a feature space.

2. Foundation models and "completeness"

The term "foundation model" first appeared in a 2021 paper titled "On the Opportunities and Risks of Foundation Models", authored by over 100 researchers associated with Stanford's "Human-Centered Artificial Intelligence" initiative (Bommasani et al. 2021). "We call these models foundation models", the paper reads, "to underscore their critically central yet incomplete character." (ibid., 1) For anyone familiar with the artificial intelligence research coming out of Stanford since the beginning of the current artificial intelligence "revolution",[1] the use of the term "incompleteness" must stand out in this definition. And in fact, "incompleteness" here needs to be understood from an entirely applied perspective. Foundation models are incomplete because they are general, and only become specific through adaptation to so-called "downstream tasks".

What sounds like humility is thus, in fact, an embrace of what has long been the implicit ideology of machine learning research: that "large" is good-enough to count as "complete", for all intents and purposes. This ideology has a surprisingly long history. Fei Fei Li, who is listed as a co-author of the paper, famously led the effort to create ImageNet, the de-facto standard for machine vision research until about 2020. ImageNet, according to Li, was supposed to "map out the entire world of objects" (Gershgorn 2017). ImageNet's ontology, in turn, builds on the ontology of WordNet – and from here we could trace the idea of capturing the world in its entirety all the way to Leibniz' and Wilkins' "universal languages" of the 17th century (the latter epitomized by Jorge Luis Borges). Back in the year 2022, the makers of Stable Diffusion, another powerful large-scale generative model, announced that their model "is the culmination of many hours of collective effort to create a single file that compresses the visual information of humanity into a few gigabytes" (Mostaque 2022).

It is thus not surprising that many humanist critiques of visual artificial intelligence[2] target this implicit ideology of completeness. It is an easy target after all, as the deficiencies are right there, to see, for everyone with the means to browse a folder full of images. That does not mean that we are discounting this kind of work. Trevor Paglen and Kate Crawford's exposure of ImageNet's wildly sexist and racist human-related categories, for instance (Crawford and Paglen 2019), or Adam Harvey's ongoing work on datasets collected "in the wild" (Harvey 2023) shifted both the academic and the public discourse towards a more critical perspective on visual artificial intelligence, even before the more recent rise of generative image models. Artistic works, such as Mimi Ọnụọha's *Library of Missing Datasets* pointed out the central role of omissions in the construction of datasets already in 2016. More recent humanist critiques of "traditional" datasets like ImageNet beyond its – now – "obvious" flaws (Birhane and Prabhu 2021) offer deep analyses of the role of the photographic apparatus in their creation (Malevé 2021), of the subtle dissonances in their categories (Smits and Wevers 2021), of their implicit ideological assumptions (Denton et al. 2021), or of the computational construction of meaning implicit in their deployment (Scheuermann et al. 2021).[3]

[1] The technical (not ideological, see above) trajectory of contemporary artificial intelligence could be traced back as far as Turing's initial considerations of machine intelligence (Turing 1950). Here, we peg the "beginning" of contemporary artificial intelligence to a paper by Krizhevsky et al. (2012), which introduces multiple groundbreaking fixes to the design and implementation of deep convolutional neural networks (CNNs) which had been around since the 1980s, including the proposal to use Web-scale data and graphics cards (GPUs) to train such networks in a massively parallel fashion.

[2] We discuss visual artificial intelligence only in this paper. For critiques – and critiques of critiques – of large language models, especially since the release of ChatGPT, see for instance Raley and Rhee 2023, Bajohr 2023, or Weatherby and Justie 2023.

[3] This is only a small slice of the critical literature on machine learning datasets. For additional examples we would like to refer the reader to the reading lists curated by the *Knowing Machines* project: https://knowingmachines.org/reading-list.

https://doi.org/10.33621/jdsr.v6i440462

And yet, the dominance of dataset critique as the default mode of academic critiques of machine learning has somewhat obscured the fact that datasets are not the only place where things go wrong. Here, we do not refer to the – equally valid and equally widely discussed – infrastructural concerns surrounding artificial intelligence in general, for instance the exploitation of "click workers" in the global south, the energy consumption of data centers, the privacy and copyright concerns connected to indiscriminate data scraping practices, or the extractivist nature of model production. Instead, our analysis – and our critique of dataset critique as a dominant practice in the humanities – focuses on the model itself.

As Sara Hooker (2021) argues, a "surprisingly sticky belief is that a machine learning model merely reflects existing algorithmic bias in the dataset and does not itself contribute to harm" while model design actually contributes just as much to a flawed final outcome. This becomes especially relevant as, with the advent of foundation models and corresponding datasets, dataset critique becomes more and more intractable as dataset sizes grow, in orders of magnitude rather than linearly. A recent example is the LAION family of datasets used to train the highly influential multimodal foundation model CLIP (Radford et al. 2021). Such datasets can only be made tractable by means of image retrieval systems. In the case of LAION, that system is, itself, powered by CLIP.[4] In fact, the most significant recent general critique of a LAION dataset (Birhane et al. 2021) uses exactly this CLIP-powered tool to identify the "misogyny, pornography, and malignant stereotypes" on which its main argument rests.

We have, then, seemingly passed a quantitative-qualitative threshold, beyond which the model itself necessarily needs to become the site of critique. Because it always played a role (as argued above), but most importantly because the simple increase in scale means that dataset critique has, or will soon, run out of steam.

3. The cultural technique of semantic compression

If ideology is now a function of the model and not (exclusively) the dataset, then our attention has to also shift towards the general cultural techniques of modeling. The term cultural technique, here, implies a set of practices that come before their theoretization. As Thomas Macho (2003) and others have argued, it is this pre-theoretical status that defines a cultural technique. So, what are the cultural techniques, the pre-theoretical (in an epistemic, not technical sense) practices that enable machine learning today? We argue that the most politically and economically relevant of these cultural techniques is semantic compression: the mapping of (representations of) the visual world to points in a feature space. It is this feature space where politically and economically salient human concepts (like race, as we discuss below) are "dissolved", and from which seemingly ahistorical, apolitical, and non-ideological versions of the same concepts are subsequently resurrected.

Semantic compression, here, denotes a form of compression that only has become possible with recent advances in machine learning. Instead of looking for formal redundancy, semantic compression algorithms, in the form of neural networks, automatically learn to represent an image "holistically" as a set of features that are "meaningful", i.e. a set of features that require gestalt knowledge of the world.

A concrete example is, again, the CLIP model. CLIP embeds both texts and images into a common embedding space. In its most prominent pre-trained manifestation – the one released by OpenAI in 2021 – its embedding space embodies the interrelations between about 400 million text and image pairs in its training data. The similarity between two texts, between two images, or between an image and a text can thus be estimated by their relative positions in this space. To illustrate this with a trivial example: a "syntactic" compression of an image of a tomato would likely latch onto the abundance of the color red in the image. "Red" would be encoded with less bytes than other colors,[5] as it has the greatest probability of appearing in the image. "Semantic" compression, on the other hand, assigns a fixed-sized vector to the

[4] https://rom1504.github.io/clip-retrieval/
[5] This is the simplest form of lossless compression, called Huffman coding.

https://doi.org/10.33621/jdsr.v6i440462

image that only becomes meaningful through its proximity to other vectors that also encode (different) images of tomatoes. The image is thus not only compressed much more efficiently (the vector space size of CLIP is 512, so each image is encoded with just 512 floating point numbers), it also becomes commensurable. This commensurability is simply a side effect in neural networks used for classification but in models like CLIP it is rendered into a powerful tool for image retrieval and many other downstream tasks. In the DALL·E 2 training pipeline, it is the CLIP model that brings together textual ("prompt") and visual information. This "encoding" step is followed by a "decoding" step facilitated by a diffusion mode (GLIDE in the case of DALL·E 2) that has learned to reconstruct legible images from Gaussian noise, again "guided"[6] by CLIP's knowledge about text-image similarity.

Obviously, semantic compression has limitations. Ted Chiang, in a 2023 New Yorker article (Chiang 2023), famously compares the large language model ChatGPT to "a blurry JPEG of all the text on the Web", implying, like so many other critics of artificial intelligence, that what we are dealing with is, essentially, nothing new. Information provided by ChatGPT – or so one could summarize Chiang's critique – is "low quality" information that is, akin to Hito Steyerl's concept of the "poor image" (Steyerl 2009), only interesting because of its technical degradation, not despite of it. The general discursive obsession with ChatGPT's "hallucinations" speaks a similar language of vaguely psychoanalytical dismissal: the knowledge space of the model is defined by a fundamental lack, a lack that can never be mitigated exactly because compression can only ever produce less, not more information.

Counterintuitively, however, this only partially holds for semantic compression. Certainly, we cannot produce something out of nothing, and the Shannon-Nyquist theorem is valid[7] no matter the scale of the dataset and no matter the complexity of the compressor. But at the same time, the feature space of contemporary image models is "filled"[8] with what can only be described as "machinic" concepts: concepts that are resurrected from the ruins of their human equivalents. As Impett and Offert show (2023), recent multimodal models can very well deal with multiple meanings, other than generative models like DALL·E seem to suggest. Concepts in feature space, in other words, are almost always more complex than any singular artifact that can be generated from them: an *in*visual culture.

4. Machinic concepts in invisual culture

In their discussion of platforms and contemporary visual culture, Adrian Mackenzie and Anna Munster (2019) use the term "invisual" to describe the changing epistemic status of images in the age of intensified computationalism. Digital images, they argue, should increasingly be understood as ensembles of computational relationalities that encompass processes like data collection and formatting, as well as systems like platforms, software, and hardware. These "ensemble images" create new orders of seeing and perception that "exceed the limits of human visual imagining" (Mackenzie and Munster, 2019: 4). The concept of the invisual gestures to what vision becomes, what images become, when their primary operations are no longer tied to once familiar humanist regimes of aesthetics, indexicality, iconicity, and representation but rather exist as artifacts of operativity. They write:

> These contemporary image ensembles are not simply quantitatively beyond our imagining but qualitatively not of the order of representation. Their operativity cannot be seen by an observing 'subject' but rather is enacted via observation events distributed throughout and across devices, hardware, human agents, and artificial networked architectures, such as deep learning networks…

[6] The exact technical process is more complex, see Ramesh et al. 2022, Radford et al. 2021.
[7] The Shannon-Nyquist sampling theorem (Shannon 1949) describes a hard limit for the amount of information that can be reconstructed from a compressed representation. See also Offert (2021).
[8] The idea of a feature space "filled" with images technically, of course, could not be further from the truth. There is nothing *in* a feature space because it is not a Euclidean space. Instead, the term only refers to the entirely artificial (and often only experimentally determined) mathematical boundaries that a compression algorithm must adhere to. "Embedding" images into a feature space requires additional algorithmic labor, as does "making" images, which also requires a whole additional algorithmic apparatus.

https://doi.org/10.33621/jdsr.v6i440462

> The visual itself as a paradigm for *how to see and observe* is being evacuated, and that space is now occupied by a different kind of perception. (Mackenzie and Munster 2019, 5 - 6)

This new kind of perception, what they call "platform seeing," is the byproduct of a visual culture that is reliant on platforms as intermediary agents. Acts of seeing digital images are not just matters of optical exposure or bodily experience but an event coordinated through the technical, economic, and political interests and affordances of these agents. This mode of perception is, paradoxically, defined by a certain imperceptibility. This is in part due to the economy of value that drives these actors and their engagement in forms of strategic opacity and trade secrecy in order to protect their own corporate interests. But it is also largely due to the impossibility of ever really knowing algorithms themselves, a pursuit that as Louise Amoore reminds us is only ever "partial, contingent, oblique, incomplete and ungrounded" (Amoore, 2020: 19). These impossible acts of imperception are what structure the invisuality of contemporary visual culture.

If we return to the example of DALL·E 2, we can observe this invisuality in action. DALL·E 2's models do not rely on semiotic or representational meaning to produce output images. Instead, they rely on cultural techniques like semantic compression and feature extraction. For instance, when a model is presented with an image of a cat it doesn't understand it as signifying a furry domestic creature that's sometimes ginger but can also be tan, gray, blue or striped, or as a Broadway musical that was devastatingly adapted into a Hollywood feature film. The model only understands it as an input with a cluster of similar, numerically valued pixels, which are statistically correlated with the text letters C A T. These correlations are created by extracting the features of images and inferring them with the features of text. But exactly how these features are identified or what even counts as a feature can be very unclear, even to the developers who program these systems. So, while DALL·E 2 may be able to create images of cats that strongly align with our own understanding of what a cat is and should look like, it doesn't do so with a coherent concept of a cat in mind; or rather, it doesn't do so using human concepts of cats in mind. Instead, it constructs invisually and through its own internal methods a cat in the image of its own machinic concepts.

A significant point we wish to impress here is that these machinic concepts are not more or less inferior, or more or less complex than their human equivalents. Our argument *is not* that computation and compression oversimplify or misinterpret human culture, as if culture is somehow separate or distinct from the technical means that produce it. Nor is it that there is a nostalgic pre-computational, pre-AI moment to which we should return. Rather, our point is that the forms of meaning-making produced through these techniques allow concepts to do different things, things that previous forms of meaning-making did not do. By using terms like "invisual" and "machinic concepts", our aim is to add to the ever-growing array of conceptual terminology responding to the shifts and changes in visual culture in the context of machine ways of seeing (Azar, Cox, and Impett 2021).

5. The durability of whiteness

What happens when these invisual, and arguably inhuman, modes of production brush against all-too-human concepts like race? (see also Phan and Wark, 2021) Rather than fixating on the problem of "solving" bias (of which there are already many efforts to do), we instead wish to focus on the durability of certain concepts even as they move through the machinations of invisual forms of image-making. In this case, the durability of whiteness. Despite the invisual abstraction of dataset images into features (supposedly value-free numerical relations), despite the encoding of concepts through processes of semantic compression, and despite the machinic forms of decoding created by opaque feature spaces, whiteness – as a structuring relation and as a tool for political erasure – continues to endure. Indeed, whiteness is so durable it can only be interrupted using blunt solutions like placing randomized, hidden keywords at the end of prompts. While much of the commentary on whiteness, bias, and generative

https://doi.org/10.33621/jdsr.v6i440462

artificial intelligence has focused on the elimination of these phenomena as fixable problems of datasets (i.e. dataset critiques), our approach is to see them instead as an exemplary moment in which models (and not data) become the new problem space and site for cultural analysis, and which have significant bearing on how we come to know significant human cultural concepts, like race.

So, what does this new problem space reveal to us about race and whiteness? As discussed above, the images that DALL·E 2 generates can be described as what Mackenzie and Munster (2019, 7) call "nonrepresentational." They are not "of the order of the visual" (ibid.) and are not contingent on semiotic meaning or humanistic ideological frames. They are not, to borrow a phrase from Lorraine Daston (2015), "epistemic images" but invisual images. Yet as the example of racial bias demonstrates, the precise problem with the images produced by DALL·E 2 is not that they are *nonrepresentational*, but rather that they are *far too* representational, relentlessly showing us the whiteness we wish we didn't have to see.

Implicit within all claims of racial bias is the assumption that there has been some kind of representational failure. That there are too many or too few racialized bodies in this or that category and that this failure extends or compounds histories of racialized misrepresentation or erasure. It is an expression of disappointment that a reality that *should be* represented *has failed to be* represented. From the perspective of dataset critique, the tendency has been to view this failure as a problem of lack (of, for example, diverse faces) that then manifests as a lack of imagination in models. If only we could fulfill the promise of "completeness" then our foundation models would no longer be lacking, and as a consequence, neither would our representational imagination. But instead of viewing bias through this prism of lack, what if we instead turn our orientation to the re-making of what is present: whiteness as a machinic concept.

In the case of DALL·E 2, the representational failure was not a question of bias (of statistical error or lack) but instead one of accuracy: the devastating reproduction of a whiteness that haunts Western visual culture. This is a criticism that has been leveled against cultural institutions like museums and galleries, institutions that have traditionally collected and assigned value to images, and to emerging custodians of new image archives, like Google images, Flickr, Wikipedia, and stock image websites whose datasets are used to train models like CLIP and subsequently DALL·E 2. While it is clear that whiteness, rather than bias, is the problem of foundation models, whiteness – as the "standard by which certain 'differences' are measured, centered and normalized" and the principle that structures hierarchies of racial domination (Moreton-Robinson 2020) – is rarely named in system cards like those used to explain the risks and limitations of the DALL·E 2 model (OpenAI 2022a).

Bias, as it turns out, is the word that is used when whiteness is too uncomfortable to say. But DALL·E 2 is not attuned to account for people's discomfort – an embodied feeling (a feeling that requires a body) that is generally expressed through silence, absence, and avoiding or moving away from the topic (see Ahmed 2012). It has no human inhibitions and so has no problem with naming and representing the whiteness that DALL·E 2 data scientists so desperately wish to avoid. What is fascinating here is that despite all efforts to keep whiteness unnamed, whiteness when resurrected as a machinic concept somehow returns in force and in ways that, ironically, helps us to see the problem of the human as a category. It helps us see (via invisual methods) that whiteness still haunts the category of human as a consistent latent feature. As outlined above, a literal unmarked sign is used as a method to address whiteness as a default category. This technique is arguably an effective way of troubling the unmarked category of the human; a byproduct that happens to align with the anti-racist threads of Black, feminist, and critical approaches to posthumanism scholarship (see for e.g. Fanon 2008; Hayles 1999; Haraway 1997, 2004; Jackson, 2020; Weheliye, 2014). At the same time, OpenAI's 'solution' to the same problem, which coincides with the commercialization of DALL·E 2, is arguably a mode of anti-racism that is primarily self-serving and which operates to conceal rather than confront racial injustice. It is a mode of anti-racism that operates in line with what Sara Ahmed (2012) describes as liberal multiculturalism's management of difference. One that bolsters ideals like "diversity" and "inclusion" as a means to enable business as usual rather than any social justice aims.

https://doi.org/10.33621/jdsr.v6i440462

Nothing demonstrates this as vividly as the most recent (early 2024) example of prompt transformation gone awry. In Google's family of Gemini multimodal models, which, at least at the time of their release, were less restricted on the prompt level than comparable models, the technique facilitated the creation of "racially diverse Nazis" (Robertson 2024), among other wildly historically inaccurate concoctions. These images – which we do not reproduce here to not give them further exposure – show exactly what it means to understand racial equity as an inverse censorship problem: it means erasing those historical injustices that have produced the problem being "fixed" in the first place. What has often been cited as an "overcorrection" is actually an "undercorrection": the prompt-transformed world, ironically, is a world in which prompt transformation is *not* necessary.

6. Discussion: Complicity and critique

Our broader claim, then, is that current-generation machine learning models require current-generation modes of (humanist) critique. It simply does not suffice anymore to point out a lack – of data, of representation, of subjectivity – in machine learning systems when these systems are designed and understood to be complete representations of reality. It is not enough anymore to simply show the whiteness we wish we did not have to see, something that popular accounts of artificial intelligence tend to do compulsively, as if one could not imagine a predominantly white and male group of doctors, engineers, or architects. Instead, we have to identify the different technical modes of whiteness at play, and understand the reconceptualization and resurrection of whiteness as a machinic concept. Feature space, in other words, needs to be understood as a political space. Here, we are following Louise Amoore's argument that for computer science all space is a potential feature space which, in turn, is always also a political space because it "can settle on what is important" (Amoore 2021, 4). We are witnessing, as she writes elsewhere "a transformation from algorithmic rules conceived to tame a turbulent, divided, and capricious world, to the productive generation of turbulence and division from which algorithmic functions are derived" (Amoore 2022, 3). In short, social fracturing is the effect of "political problems becoming reconfigured as machine learning problems" (Amoore 2022, 3).

At the same time, we have a less-than-steady grasp on feature space. One basic example is the insufficiency of visualizations, for instance cluster visualizations like those provided by PixPlot (for images) or Google's Embedding Projector (for text). As Peter Galison (2002) has argued, we have always needed images to make sense of the world. At the same time, the simple presence of images often obfuscates their inadequacy as vehicles of representation. We should not rely on them too much or hope they will speak the truth. Without images, however, we "cannot proceed further towards abstraction," (ibid.) simply because we cannot think in purely symbolic operations. Problematically, high-dimensional vector spaces are geometrically counter-intuitive spaces. Not only is it impossible to imagine a vector in 512 dimensions, high dimensional vector spaces have counter-intuitive properties as well: this is what is known as the "curse of dimensionality". "Under certain broad conditions […], as dimensionality increases, the distance to the nearest neighbor approaches the distance to the farthest neighbor" (Beyer et al. 1999). In other words: in latent space, distances between data points tend to become illegible. They lose, or at least significantly change their meaning and thus become inaccessible to intuition. Accordingly, a mapping-back from latent space to Euclidean space (for instance via T-SNE, U-MAP, or simple PCA) can again only be "lossy", both mathematically and epistemically. The ridiculousness of recommendations produced by state-of-the-art industry recommender systems (for instance, buying a highly specific item, e.g. a drill press, only to be recommended to buy another drill press the next day) is a direct consequence of this limitation. In general, semantic compression has been understood as "lossy" and thus as deficient.

The takeaway, then, is neither "CS needs to have ethics" nor "humanists need to learn to code" but rather finding new experimental methodologies to interrogate feature space: this includes better theory (such as in Munster and MacKenzie), but also a closer look at the epistemic affordances of those

https://doi.org/10.33621/jdsr.v6i440462

techniques that the technical disciplines have to offer, in particular fields like interpretable machine learning (see Offert and Bell 2021) and representation learning. We have to re-evaluate the situatedness of humanist critique, and more specifically, the implicit complicity of humanist critique with the completeness paradigm. To summon Latour again: "[A] certain form of critical spirit has sent us down the wrong path, encouraging us to fight the wrong enemies and, worst of all, to be considered as friends by the wrong sort of allies because of a little mistake in the definition of its main target." (Latour 2004) More often than not, dataset critique also implies "delivering", or at least listing and suggesting for inclusion those "diverse" images that are missing from mainstream datasets.

Dataset critique, somewhat polemically speaking, can thus actually reinforce the ideological leap from "good enough" to "complete" dictated by computer science. And we are stuck with this complicity, too – because what else is there to do, without proper access, resources, and training? The current shift towards foundation models, then, at the very least presents an opportunity to reflect on what is next, even if it is just a "new and better" kind of complicity.

Author contributions

Both authors contributed equally to this publication.

References

Ahmed, S. (2012) *On being included: Racism and diversity in institutional life*, Duke University Press.

Amoore, L. (2022) 'Machine learning political orders', *Review of International Studies* 49(1), pp. 20-36. https://doi.org/10.1017/S0260210522000031

Amoore, L. (2021) 'The deep border', *Political Geography*. https://doi.org/10.1016/j.polgeo.2021.102547

Azar, M., Cox, G. and Impett, L. (2021) 'Introduction: Ways of machine seeing', *AI & Society* 36, pp. 1093-1104. https://doi.org/10.1007/s00146-020-01124-6

Bajohr, H. (2023.) 'Dumb meaning: Machine learning and artificial semantics', *IMAGE* 37(1), pp. 58-70. https://doi.org/10.1453/1614-0885-1-2023-15452

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999), 'When is "nearest neighbor" meaningful?', *ICDT'99: 7th International Conference.* Jerusalem, Israel, January 10–12, 1999, pp. 217-235.

Birhane, A., Prabhu, V.U. and Kahembwe, E. (2021), 'Multimodal datasets: Misogyny, pornography, and malignant stereotypes', arXiv preprint 2110.01963. https://arxiv.org/pdf/2110.01963.pdf

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S et al. (2021) 'On the opportunities and risks of foundation models', arXiv preprint 2108.07258. https://arxiv.org/pdf/2108.07258.pdf

Chiang, T. (2023) 'ChatGPT is a blurry JPEG of the web', *The New Yorker*, Feb. 9. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

Crawford, K., Paglen, T. (2019) 'Excavating AI: The politics of training sets for machine learning'. https://excavating.ai

Daston, L. (2015) 'Epistemic images', in Payne, A. (ed.) *Vision and its instruments: Art, science, and technology in early modern Europe*, The Pennsylvania State University Press, pp. 13–35.

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021) 'On the genealogy of machine learning datasets: A critical history of ImageNet', *Big Data & Society* 8(2). https://doi.org/10.1177/20539517211035955

Fanon, F. (2008) *Black skin, white masks*. Grove Press.

Galison, P. (2002) 'Images scatter into data, data gather into images', in Latour, B., Weibel, P. (eds.) *Iconoclash: Beyond the image Wars in science, religion, and art*, MIT Press.

Gershgorn, D. (2017) 'The data that transformed AI research – and possibly the world', *Quartz*, July 26. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world

Haraway, D. (2004) 'A manifesto for cyborgs: Science, technology and socialist-feminism in the 1980s', in *The Haraway Reader*, Routledge, pp. 7–47.

Haraway, D. (1997) *Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse: Feminism and Technoscience*, Routledge.

Harvey, A., LaPlace, J. (2021) *Exposing.ai*. https://exposing.ai/

Hayles, N.K. (1999) *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*, University of Chicago Press.

Hooker, S. (2021) 'Moving beyond "algorithmic bias is a data problem"', *Patterns* 2(4). https://doi.org/10.1016/j.patter.2021.100241

Impett, L., Offert, F. (2023) 'There is a digital art history' *Visual Resources* 38(2). https://doi.org/10.1080/01973762.2024.2362466
Birhane, A. and Prabhu, V.U. (2021) 'Large image datasets: A pyrrhic win for computer vision?', *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536-1546.
Ross, J., Irani, L., Silberman, M.S., Zaldivar, A. and Tomlinson, B. (2010) 'Who are the crowdworkers? Shifting demographics in Mechanical Turk', *CHI'10 extended abstracts on human factors in computing systems*, pp. 2863-2872.
Jackson, Z.I. (2020) *Becoming human: Matter and meaning in an antiblack world*. NYU Press.
Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet classification with deep convolutional neural networks', *Advances in neural information processing systems* 25. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
Latour, B. (2004) 'Why has critique run out of steam? From matters of fact to matters of concern', *Critical Inquiry* 30(2), pp. 225-248.
Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. (2023) 'Stable bias: Analyzing societal representations in diffusion models', arXiv preprint 2303.11408. https://arxiv.org/abs/2303.11408
Macho, T. (2003) ‚Zeit und Zahl. Kalender und Zeitrechnung als Kulturtechniken' in Krämer S., Bredekamp, H. (eds.), *Bild – Schrift – Zahl*, Wilhelm Fink, pp. 179-192. http://www.thomasmacho.de/index.php?id=zeit-und-zahl
MacKenzie, A., Munster, A. (2019) 'Platform seeing: Image ensembles and their invisualities', *Theory, Culture & Society* 36, pp. 3-22. https://doi.org/10.1177/0263276419847508
Malevé, N. (2021) 'On the data set's ruins', *AI & Society* 36(4), pp.1117-1131. https://doi.org/10.1007/s00146-020-01093-w
Millière, R. (2022) 'Adversarial attacks on image generation with made-up words', arXiv preprint 2208.04135. https://arxiv.org/pdf/2208.04135.pdf
Moreton-Robinson, A. (2021) *Talkin'up to the white woman: Indigenous women and feminism*, University of Minnesota Press.
Mostaque, E. (2022) *Stable Diffusion public release*. https://stability.ai/blog/stable-diffusion-public-release
Offert, F. (2021) 'Latent deep space: Generative adversarial networks (GANs) in the sciences', *Media+Environment* 3(2). https://mediaenviron.org/article/29905-latent-deep-space-generative-adversarial-networks-gans-in-the-sciences
Offert, F., Bell, P. (2021) 'Perceptual bias and technical metapictures. Critical machine vision as a humanities challenge', *AI & Society* 36, pp. 1133–1144. https://doi.org/10.1007/s00146-020-01058-z
Offert, F. (2023) 'On the concept of history (in foundation models)', *IMAGE* 37(1), pp. 121-134. https://doi.org/10.1453/1614-0885-1-2023-15462
OpenAI (2022a) *DALL·E 2 preview – Risks and limitations*. https://github.com/openai/dalle-2-preview/blob/main/system-card.md
OpenAI (2022b) *Reducing bias and improving safety in DALL·E 2*. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/
OpenAI (2023) *DALL·E 3 system card*. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf
Phan, T. and Wark, S. (2021) 'What personalisation can do for you! Or: How to do racial discrimination without 'race'', *Culture Machine* 20, pp. 1-29.
Raley, R. and Rhee, J. (2023) 'Critical AI: A field in formation', *American Literature* 95(2).
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal et al. (2021) 'Learning transferable visual models from natural language supervision', *International Conference on Machine Learning (ICML),* 8748–8763. https://proceedings.mlr.press/v139/radford21a/radford21a.pdf
Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022) 'Hierarchical text-conditional image generation with CLIP latents', arXiv preprint 2204.06125. https://arxiv.org/abs/2204.06125
Robertson, A. (2024) 'Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis', *The Verge*, Feb. 21. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical
Salvaggio, E. (2023) 'Shining a light on shadow prompting', *Tech Policy Press*, Oct. 19. https://www.techpolicy.press/shining-a-light-on-shadow-prompting/
Scheuerman, M.K., Hanna, A. and Denton, E. (2021) 'Do datasets have politics? Disciplinary values in computer vision dataset development', *Proceedings of the ACM on human-computer interaction* 5(CSCW2), pp. 1-37. https://doi.org/10.1145/3476058
Shannon, C.E. (1949) 'Communication in the presence of noise', *Proceedings of the IRE* 37(1), pp.10-21.
Smits, T. and Wevers, M. (2022) 'The agency of computer vision models as optical instruments', *Visual Communication* 21(2), pp.329-349. https://doi.org/10.1177/1470357221992097
Turing, A.M. (1950) 'Computing machinery and intelligence', *Mind* 59(236).
Weatherby, L. and Justie, B. (2022) 'Indexical AI', *Critical Inquiry* 48(2), pp. 381-415.
Weheliye, A.G. (2014) *Habeas viscus: Racializing assemblages, biopolitics, and black feminist theories of the human*, Duke University Press.

