Fabian Offert

On the Concept of History (in Foundation Models)

Forthcoming in: *Thinking with AI*, ed. Hannes Bajohr, Open Humanities Press, 2024

I

Any sufficiently complex technical object that exists in time has, in a sense, a concept of history: a particular way that the past continues to exist for it, with contingencies and omissions specific to its place and role in the world. Computation is no exception to this, and indeed takes its very efficacy from a particular technical relation to the passing of time. Meanwhile, the emergence of so-called 'foundation models' (Bommasani et al. 2021), a specific class of technical objects that have come to dominate the field of artificial intelligence, promises to significantly change what it means to 'compute' in the first place (Offert 2023a), and especially, I will argue, what it means to compute the past. This essay thus asks: what is the concept of history that emerges from foundation models, and particularly from large visual models? Do such models conceptualise the past? What is the past *for them*?

My question does not imply any intentionality (Searle 1980), agency, or subjectivity – real or fictional – on the part of the models under investigation. It is exactly not 'what is it like to be' a foundation model, to paraphrase Thomas Nagel (1974). The question, in other words, is entirely non-philosophical and non-speculative. It is also separate from the question of the historicity of foundation models themselves, that is, their role in a larger history of artificial intelligence, both as a general problem starting in the 1950s (see Pasquinelli 2023, Dobson 2023) and as a specific set of technical approaches that first emerged around 2012 (see Offert 2022). What remains, then, is a technical object (that is – again – certainly a product of history), or rather a class of technical objects, which, in some sense that we would need to determine, relate to the passing of time in non-trivial, non-arbitrary ways. We thus need to take a closer look at the material basis of foundation models, to trace, at all, or at least at some, levels of the stack[i] how 'history is made', that is, where exactly such non-trivial, non-arbitrary ways to deal with the passing of time emerge.

II

The very definition of the computability of a number, as first proposed by Alan Turing (1937), is that the number can be produced by a discrete state machine – a machine that moves through a finite set of deterministic configurations over time. Foundation models are computer programs, and thus take part in this necessary relation to the passing of time.[ii] But more importantly, foundation models are machine learning models, and it is the learning[iii] part where a difference in their relation to the passing of time emerges.

Consider a simple computer vision classifier, for instance a deep convolutional neural network like VGG19. 'Training' a VGG19 model means tuning its parameters, or weights, according to a dataset of images. How exactly the parameters are arranged and interconnected is what defines the architecture of the model. The parameters, in turn, define how the images are passed on through the network, and thus which predefined category they are eventually attributed to. Over the course of the whole training process, the model is exposed to millions of

images, one image at a time, and its parameters are adjusted at each step.[iv] As the parameters are usually initialised with random numbers,[v] the first steps of the training process often require large adjustments which then become progressively smaller.[vi] The model thus begins its 'life' as a somewhat malleable structure but becomes more rigid the closer it moves towards 'convergence', that is, towards a state in which it sufficiently models the inherent probability distribution of the dataset of images. From there on, the model is usually used for inference only, remaining completely unchanged for the rest of its 'life'.

Of course, inference is still a computational process, and thus on the lower levels of the stack time goes on. On the level of the model, however, it comes to a standstill, and all its history is erased. Indeed, every step of the training process is destructive by default, as parameters are irreversibly altered after each backwards pass[vii]. There is thus simply no going back to earlier points in the training process, unless they are intentionally, and separately, recorded as so-called 'checkpoints'.[viii] From looking at a fully trained model, we simply cannot tell what it 'went through', for instance how good or bad it *used to be* at its respective task. One consequence of this opaque relation, or rather non-relation, of the model to its own past is that it cannot be easily adapted to other tasks, as Christina Vagt points out in her analysis of 'catastrophic forgetting' in this volume. Another consequence is that a fully trained model cannot be understood anymore in terms of its functional 'parts'.

In fact, we could describe the training of our model as a process of concretisation in the sense of philosopher Gilbert Simondon (2016). 'Concretization', for Simondon, describes the evolution of technical objects from a state of 'functional indeterminacy' (abstract, all parts have their own internal logic) to a state of functional completeness (concrete, all side effects become synergies). And indeed, neural networks could be described as moving from an abstract state (an empty computation graph) to a concrete state (a fully trained neural network of weights). This perspective is supported technically: 'knowledge' in neural networks is always distributed,[ix] it is represented by the network as a whole, rather than by individual neurons. For Simondon the difference between the 'science' and the 'technology' of a machine is the margin of concretisation still obtainable. In the training of our model, the stated goal is to reduce this margin to zero, even if this usually turns out to be impossible in practice. Accordingly, while empty neural networks can be described technologically (e.g. in code), fully trained neural networks can only be probed empirically – what the training process leaves behind are only monuments, not documents.[x] Monuments require interpretation – and neural networks are no exception, as the question of explainable artificial intelligence and the rise of mechanistic interpretability demonstrate.

It is when we look at foundation models' reliance on data, however, that their complicated relation to the past attains special significance. There are a few relevant aspects of data that we can simply name here, as others have looked at them in great detail.[xi] Datasets emerge from processes of selection and exclusion. They do not even reflect a particular, biased view on the world but a particular, biased view on only that part of the world that is readily available in digital form. Their assembly often relies on exploitative practices and questionable interpretations of privacy and copyright. They are often based on rigid ontologies[xii] and the idea that the world can be neatly categorised without residue, a problem that goes back as far as Wilkins and Leibniz.

III

For these and many other reasons, foundation models will never facilitate anything even close to a human concept of history, one that relies on an intersubjectively negotiated, comprehensively factual, deeply archival, and necessarily causal perspective on the world – as much is clear. And yet, if we look at the recent *output* of large visual models, what we see can intuitively only be understood as 'historical'. Other than the deep relation of artificial intelligence research and science fiction would suggest, foundation models are not at all utilised to imagine the future, but to reimagine the past. One particularly striking example are stills from fictional movies – fictional as in never made[xiii] – which manage to capture the particular aesthetics of specific directors and time periods, and instill a peculiar sense of nostalgia, as Roland Meyer (2023) has argued.

Given this fixation on history in the use of foundation models, and given that, with Simondon (and mechanistic interpretability), we can only study such models empirically, our initial question should be rephrased as follows: as far as can be shown, is there a degree of consistency to the outputs of a foundation model when it is tasked with processing inputs related to the past that would suggest a model-specific 'concept of history'? And if so, what are the structuring principles of these internally consistent outputs, and how do they relate to the structuring principles humans apply to the past to render it history? Or, as this essay focuses on visual models: what happens to human visual culture when it is processed by a foundation model if visual culture is indeed 'what is seen', and if 'what is seen' is indeed 'what changes over time' (Roeder 1988, quoting Gertrude Stein)?

My experimental close readings of one such system in particular, the CLIP model released by OpenAI in 2021, suggests that one of these structuring principles, and arguably the most significant at least for visual models, is a technically determined form of *remediation* (Bolter and Grusin 2000). Polemically, for CLIP and CLIP-dependent generative models like DALL·E 2, the recent past is literally black and white, and the distant past is actually made of marble. Given that CLIP, at the same time, *premediates* our future digital experience as a means of search, retrieval, and recommendation, this structuring principle of remediation then becomes ethically and politically relevant. As Alan Liu asks:

> Today, the media question affects the sense of history to the core. [...] This is not just an abstract existential issue. It's ethical, political, and in other ways critical, too. Have we chosen the best way to speak the sense of history today, and if so, for the benefit of whom? (Liu 2018: 2)

The ethical questions surrounding this 'media question' are maybe nowhere as obvious as in the digitisation of the testimonies of those who survived the Holocaust (Walden and Marrison 2023). Projects like *Dimensions in Testimony*, which is funded by the USC Shoah Foundation, have started to go beyond the mere recording of testimonies, attempting to emulate their performative quality, the significant experience of sharing a moment in space and time, with the help of artificial intelligence. As the project website states:

> Dimensions in Testimony enables people to ask questions that prompt real-time responses from pre-recorded video interviews with Holocaust survivors and other

> witnesses to genocide. The pioneering project integrates advanced filming techniques, specialized display technologies and next generation natural language processing to create an interactive biography. (USC Shoah Foundation, 2023)

Todd Presner (2022) has pointed out the dilemma that such projects find themselves in. In *Dimensions in Testimony,* he argues, humans 'are no longer (centrally) part of the creation of digital cultural memory'. Instead, through established and artificial intelligence-enhanced technologies of montage, individual testimonies, once irreversibly tied to an individual human life, become disembodied. If the duty to keep these testimonies accessible for future generations warrants these technological interventions – 'that Auschwitz not happen again',[xiv] in Adorno's words – is an open question. Irrespective of such ethical considerations, projects like *Dimensions in Testimony* point to a fundamental media-theoretical question about the ethics of memory, and, by extension, the concept of history: What is the imprint that a specific technology[xv] leaves on history? More precisely, what, if anything, do foundation models 'add' to an already (re-)mediated past?

IV

Here, we need to turn to Walter Benjamin's text *On the Concept of History* (Benjamin 2006a) that the title of this essay takes inspiration from. Years of scholarly debate on Benjamin's writings in general, and his concept of history in particular,[xvi] have made it unnecessary to introduce its premise here, or comment on the unusual synthesis of materialist and theological thought that it embodies. Instead, I would like to point out an almost trivial similarity between *On the Concept of History* and Benjamin's other widely read essay on the *Work of Art in the Age of Its Technological Reproducibility* (Benjamin 2006b).

Famously, in *On the Concept of History,* Benjamin writes: 'Articulating the past historically does not mean to recognise it "the way it really was" […]. It means appropriating a memory as it flashes up at a moment of danger'.[xvii] (391) Previously, in the *Work of Art* essay, Benjamin had argued that the political potential of film derives from its potential to produce abrupt cuts, and thus 'shock' (267) the viewer into a different mode of thinking. In other words, for Benjamin, the condition under which history becomes possible, the 'moment of danger' is the condition that film emulates. In both cases, awareness and insight depend on a moment of immediacy, and in both cases this moment of immediacy must be actively captured and repurposed for a progressive (Marxist) agenda before it falls into the hands of the fascists. There is thus, for Benjamin, a structural similarity between history as a memory that 'flashes up', that emerges from, and is actualised by, a moment of crisis, and the specific ways in which technology mediates our experience of the present world, and thus shapes our political views of it. Crucially, history and technology manifest themselves as a specific way of seeing.

What I am suggesting here, then, is not that we should 'apply' Benjamin's concept of history to artificial intelligence systems. On the contrary: One of the reasons why the field of 'critical AI studies'[xviii] has not had the impact that one would expect given the oversized importance of artificial intelligence research in computer science, is its insistence on resorting to traditional humanist theoretical frameworks and concepts that simply do not suffice anymore. Instead, I would like to propose, exactly with Benjamin, that we have to carve out the extremely specific, borderline idiosyncratic ways of seeing that artificial intelligence systems bring to the

table where they are tasked with processing, or producing, an already mediated past. Again, more precisely: as the past is remediated through contemporary artificial intelligence systems, is the concept of history that emerges from this process of remediation different from the concept of history that emerges from the always already (re-)mediated data on its own? What, in other words, is the surplus remediation inherent in a foundation model's specific way of seeing? These questions also bring us back to the title of this volume. 'Thinking with AI', in this context, means to understand artificial intelligence as an opportunity to re-think which levels of the stack a humanist analysis of computation needs to address to be of critical value.

'Foundation model' is a term introduced by a collective of researchers at the Stanford HAI institute in 2021 (Bommasani et al. 2021). It basically means models that are a) very large, and b) that can be used for a variety of 'downstream' tasks. The vision model CLIP (Contrastive Language-Image Pre-Training), first released in 2021(Radford et al. 2021) by OpenAI, is such a foundation model. Outside the technical community, its innovations were somewhat obscured by the concurrent release of the DALL·E model, and later overshadowed by DALL·E's successor, DALL·E 2 (Ramesh et al. 2022) and the language model GPT-3.

CLIP – other than both iterations of DALL·E, as well as GPT-3 – is not a generative model. It does not produce images or text, but it connects them. More precisely, CLIP learns from images in context by projecting an image and its context into a common 'embedding space'. The 'context' here could be an image caption, a so-called 'alt text' which describes the image in case it is not loaded properly and to accommodate people with screen readers, or simply a news article that the image illustrates. A fully trained CLIP model, then, consists of a high-dimensional vector space, or embedding space, in which words and images that are related can be found close together. Similarity between image and text is thus modeled as spatial proximity (this is true for all embedding models, be it just words, just images, or both, such as in the case of CLIP). While CLIP was originally designed for zero-shot image labeling,[xix] it also facilitates what computer scientists call 'image retrieval' (this exemplifies its 'foundation' character): finding specific images within an unlabeled corpus of images based on visual or textual prompts. The user can provide CLIP with an image and it will look for similar images, or they can provide it with a prompt and it will look for images corresponding to this prompt – in any corpus of images. Given that the training corpus for CLIP is largely unknown,[xx] it seems futile to attempt to construct a somewhat empirical basis for our claims. And yet, there are two ways to study CLIP's concept of history empirically

VI

The first way we could call 'attribution by proxy'. While we do not know what CLIP was trained on, we can still ask it for things *in terms* of specific collections of images. It is exactly this aspect of CLIP – the universality of its embeddings – that makes it so powerful as a retrieval engine. The following examples were produced with a custom CLIP-based search engine called imgs.ai (Offert and Bell 2023), which indexes museum collections in the public domain.

To illustrate the conceptual depth of CLIP, consider the search prompt 'rhythm', applied to the (digitised) collection of Museum of Modern Art, New York, which contains about 70,000 images in total. If we query the collection with this (intentionally abstract) prompt, we will receive a selection of images which reflect the polyvalence of 'rhythm': images of sheet music, album covers, and loudspeakers, works that resemble oscilloscope graphs or spectral plots, or

graphical works that involve regular patterns that could be described as ‘rhythmic’.

Going back to the ethical and political stakes of automated vision, we can query this same collection for ‘images of the Holocaust’. And the results tell us that, yes, CLIP knows – too well – what we are talking about. On the one hand, the model will suggest those few images in the MoMA collection that are historically linked to the query, for instance photographs by the U.S. Army Signal Corps which played an important role in documenting the atrocities of Nazi Germany. But on the other hand, it will exemplify a much more abstract knowledge about visual Holocaust memory. Suggested results include a photograph by Bruce Davidson, shot on the set of the war film *Lost Command* in Spain in the 1960s,[xxi] a 1980 photograph by Aaron Siskind depicting volcanic lava,[xxii] or a collage made from stamps by Robert Watts in 1963.[xxiii] None of these pictures are historically related to the Holocaust, nor are they necessarily meant to evoke it, but all of them could be easily recontextualised with respect to the visual language of Holocaust cultural memory. Using the MoMA collection as a proxy, we can see how well CLIP has internalised this visual language. Moreover, far from just showing the unshowable, CLIP has clearly learned that this language operates metaphorically. But: the fact that all the results that CLIP proposes (not only those named above) are black-and-white photos already points to a significant limitation, a limitation that we can further explore by utilising generative models. This second way of studying CLIP we could call ‘generative attribution’. It is made possible by the fact that CLIP, to a large part, determines the training of generative models like DALL·E and Stable Diffusion.

VII

If we ask DALL·E 2 for ‘a color photo of a fascist parade, 1935’ it will not comply. ‘Fascism,’ among many other political terms, was banned by OpenAI, early on, to mitigate the potential of their model – of which they were well aware – to produce politically, legally, or socially unacceptable material like deep fakes, pornography, or propaganda. Such safeguards are not in place in other models like Stable Diffusion but there exists a simple trick to circumvent DALL·E’s forced ‘neutrality’ as well. Intentionally misspelling ‘fascism’ by leaving out the ‘s’[xxiv] will produce (a variation of) the image in figure 1: a vaguely Western European city with some sort of mass rally taking place, red flags raised, and ominous smoke emerging from a building in the background. DALL·E, in other words, despite its safeguards, knows very well what 1935 fascism looks like – *to us.* The generated image has the appearance of a historical photograph not only for its subject but for its appearance; it shows the characteristic colours of early Kodachrome slide photography, with the red of the flags particularly standing out against an otherwise subdued sepia palette. This is how Nazi Germany appears in the photographs of Hugo Jäger, for instance, whose pre-war slide collection was acquired and popularised by *LIFE* magazine in the 1960s.[xxv]

Figure 1. DALL·E 2 generation for 'a color photo of a facist [sic] parade, 1935', produced in October 2022. Note that this safeguard circumvention technique has been 'fixed' at the time of writing.

What is remarkable about this generated image is not its accuracy in emulating a specific historical medium – this has been possible at least since the early days of style transfer ca. 2016 – but that it resorts to this specific historical medium by default. Nowhere in the prompt did we ask for early Kodachrome in particular. And it turns out that it is hard to get rid of, too. From experiments done on both DALL·E 2 and Stable Diffusion, it is difficult to impossible to produce colour photographs of fascist parades, ca. 1935, that do *not* have the appearance of early Kodachrome, colourised black-and-white, or otherwise historically more or less accurate photographic techniques. Only through copious amounts of highly specific additional keywords or negative prompts – prompts which explicitly describe which kind of outputs should be avoided – is it possible to steer the model away from this particular aesthetic. There exists, in other words, a strong default in models like DALL·E that conjoins historical periods and historical media and thus produces a (visual) world in which fascism can simply not return

because it is safely confined to a black-and-white (or, in our case, Kodachrome) media prison.

VIII

Of course, all of this is, in a way, not very surprising. Before the invention of photography, history was not associated with black-and-white at all. The past, in other words, for us and the model, exists visually only through those historical media that we see emulated here. 'Media determine our situation' (Kittler 1999: xxxix), for better or worse, and it is hard for us, too, to picture the past alive. And yet, the current generation of foundation models can easily produce highly speculative images when the speculation concerns the content, not the style, of the image. Contemporary generative models are famously able to generate entirely fictional images like the well-known 'astronaut riding a horse on the moon'. While DALL·E 2, for instance, has no problem producing a cartoon image of a cat driving a car, a realistic colour photograph of a cat driving a car – where the cat actually drives the car, paws on the steering wheel – again requires copious amounts of prompt engineering.

Figure 2. DALL·E 2 generations for 'Laocoön and his sons, between 27 BC and 68 AD' and 'Tank Man, 1989', both produced in October 2022.

The flip side of this capability is that it cannot be switched off easily. In the case of proprietary models like DALL·E 2, which includes additional safeguards that are supposed to guarantee it remains 'culturally agnostic' (Cetinic 2022), this has significant consequences. While 'allowed', *generally* historical prompts (including those originally hidden behind surface-level, that is, prompt parsing safeguards, like 'fascism') are tied to specific forms of mediation, *specifically* historical prompts are decoupled from the event that they refer to and relegated to a world of fiction. Why? Because the model *must have an answer*. As for all foundation models, failure is not an option – there has to be *a* result, no matter how outrageous. Foundation models, in other words, are *contingency machines*.[xxvi] DALL·E 2, in particular, fails to reproduce historical images without altering their meaning. The prompt 'Laocoön and His Sons, between 27 BC and 68 AD' which references the famous work central to European art history since

Winkelmann, produces a serene image of a Black[xxvii] family with no trace of agony. The prompt ‘Tank Man, 1989’, which references the iconic photograph from the Chinese Tiananmen protests, produces an image of a soldier proudly looking at a tank, rather than a scene of radical civil disobedience (both figure 2).

IX

What, if anything, does artificial intelligence ‘add’ to an already mediated past? We now have to state that artificial intelligence not only adds nothing, but it forecloses a political potential. Models like DALL·E 2 find themselves in a triple bind: they suffer from syntactic invariability in the case of *generally* historical prompts, semantic arbitrarity in the case of *specific* historical prompts, and superficial, corporate censorship that affects both. The result is an implicitly politicised concept of history. In the most literal interpretation of the famous idea that history doesn’t repeat itself, the past can never be actualised and is eternally tied to a specific medium, while images that are already rendered into history are excluded from making an appearance by simple corporate policy. Neither can history be made by actualising the past for the present, nor can the already-historical past be summoned. One of the many consequences is a (visual) world in which fascism can simply not return because it is, paradoxically at the same time, censored (we cannot talk about it), remediated (it is safely confined to a black-and-white media prison), and erased (from the historical record).

Works Cited

Aaronson, Scott. 2013. *Quantum Computing since Democritus*. Cambridge University Press.

Adorno, Theodor W. 1970. 'Erziehung nach Auschwitz.' In *Erziehung zur Mündigkeit: Vorträge und Gespräche mit Hellmuth Becker 1959-1969*, edited by Gerd Kadelbach, pp. 135-162. Frankfurt am Main: Suhrkamp.

Barthes, Roland. 1982. 'The Reality Effect.' In *French Literary Theory Today: A Reader*, edited by Tzvetan Todorov, pp. 11-17. Cambridge University Press.

Benjamin, Walter. 1974. 'Über den Begriff der Geschichte.' In *Gesammelte Schriften I.2*, pp. 693-704. Frankfurt am Main: Suhrkamp.

Benjamin, Walter. 1974b. 'Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit.' In *Gesammelte Schriften I.2*, pp. 471-508. Frankfurt am Main: Suhrkamp.

Benjamin, Walter. 2006a. 'On the Concept of History'. In *Selected Writings*, vol. 4, edited by Michael W. Jennings. Harvard University Press.

Benjamin, Walter. 2006b. 'The Work of Art in the Age of Its Technological Reproducibility'. In *Selected Writings*, vol. 4, edited by Michael W. Jennings. Harvard University Press.

Bolter, Jay David and Richard Grusin. 2000. *Remediation: Understanding New Media*. MIT Press.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. 2021. 'On the Opportunities and Risks of Foundation Models.' arXiv preprint 2108.07258.

Cetinic, Eva. 2022. 'Multimodal Models as Cultural Snapshots.' Talk given at Ludwig Forum Aachen, November 18.

Cherti, Mehdi, et al. 2022. 'Reproducible Scaling Laws for Contrastive Language-Image Learning.' arXiv preprint 2212.07143.

Cosgrove, Ben. No date. 'A Brutal Pageantry: The Third Reich's Myth-Making Machinery, in Color.' *LIFE History*. https://www.life.com/history/a-brutal-pageantry-the-third-reichs-myth-making-machinery-in-color.

Didi-Huberman, Georges. 2017. *The Surviving Image: Phantoms of Time and Time of Phantoms: Aby Warburg's History of Art*. Pennsylvania State University Press.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. MIT Press.

Dobson, James. 2023. *The Birth of Computer Vision*. University of Minnesota Press.

Impett, Leonardo, and Fabian Offert. 2023. 'There Is a Digital Art History.' arXiv preprint 2308.07464.

Kittler, Friedrich. 1990. *Discourse Networks 1800/1900*. Stanford University Press.

Kittler, Friedrich. 1999. *Gramophone, Film, Typewriter*. Stanford University Press.

Kittler, Friedrich. 2012. 'There Is No Software'. In: Literature, Media, Information Systems. New York, NY: Routledge.

Krämer, Sybille. 2006. 'The Cultural Techniques of Time Axis Manipulation: On Friedrich Kittler's Conception of Media.' *Theory, Culture & Society* 23, no. 7-8: 93-109.

Liu, Alan. 2018. *Friending the Past: The Sense of History in the Digital Age*. University of Chicago Press.

Löwy, Michael. 2005. *Fire Alarm: Reading Walter Benjamin's 'On the Concept of History'*. London: Verso.

Meyer, Roland. ‘Die Nostalgiemaschine’. *54books*. Forthcoming.
Nagel, Thomas. 1974. ‘What is it like to be a bat?’ *The Philosophical Review* 83, no. 4: 435-450.
Offert, Fabian and Thao Phan. ‘A Sign That Spells: DALL·E 2, Invisual Images and The Racial Politics of Feature Space.’ arXiv preprint 2211.06323.
Offert, Fabian. 2023a. ‘Can We Read Neural Networks? Epistemic Implications of Two Historical Computer Science Papers.’ *American Literature* 95, no. 2.
Offert, Fabian and Peter Bell. 2023. ‘imgs.ai. A Deep Visual Search Engine for Digital Art History.’ *International Journal for Digital Art History*. Forthcoming.
Offert, Fabian. 2023b. ‘On the Emergence of General Computation from Artificial Intelligence.’ https://zentralwerkstatt.org/blog/on-the-emergence-of-general-computation-from-artificial-intelligence.
Offert, Fabian. 2022. ‘Ten Years of Image Synthesis.’ https://zentralwerkstatt.org/blog/ten-years-of-image-synthesis.
Panofsky, Erwin. 1955. ‘The History of Art as a Humanistic Discipline’. In: *Meaning in the Visual Arts*. University of Chicago Press.
Pasquinelli, Matteo. 2023. *The Eye of the Master. A Social History of Artificial Intelligence*. London: Verso. Forthcoming.
Pavich, Frank. 2023. ‘This Film Does Not Exist.’ *New York Times*. January 13, 2023.
Presner, Todd. 2022. ‘Digitizing, Remediating, Remixing, and Reinterpreting Holocaust Memory.’ Talk given at the University of California, Santa Barbara, May 10.
Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. 2021. ‘Learning Transferable Visual Models from Natural Language Supervision.’ *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139.
Raley, Rita and Jennifer Rhee. 2023. ‘Critical AI: A Field in Formation.’ *American Literature* 95, no. 2.
Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. ‘Hierarchical Text-Conditional Image Generation with CLIP Latents.’ arXiv preprint 2204.06125.
Roeder, George H. Jr. 1988. ‘Filling in the Picture: Visual Culture.’ *Reviews in American History* 26, no. 1 (March): 275-293.
Simondon, Gilbert. 2016. *On the Mode of Existence of Technical Objects*. University of Minnesota Press.
Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. ‘Intriguing properties of neural networks.’ arXiv preprint 1312.6199.
Turing, Alan. 1937. ‘On Computable Numbers with an Application to the Entscheidungsproblem.’ *Proceedings of the London Mathematical Society* s2-42, no. 1: 230-265.
USC Shoah Foundation. 2023. *Dimensions in Testimony*. https://sfi.usc.edu/dit
Walden, Victoria Grace and Kate Marrison. 2023. ‘Recommendations for Digitally Recording, Recirculating, and Remixing Holocaust Testimony: Digital Holocaust Memory Project Report.’ Sussex Weidenfeld Institute of Jewish Studies.

[i] The term ‘stack’ is used here in a precise technical, rather than a philosophical sense to facilitate what Leonardo Impett has called ‘full stack critique’: identifying the epistemic and, by extension, political implications of the concrete technical decisions from which a technical object emerges. Underlying this is the assumption that such implications are indeed distinct, and cannot be collapsed into the material realm, as Friedrich Kittler (2012) has argued.

[ii] There remains much to be said about the peculiar relation of computation to Kant’s two pure forms of intuition. On the one hand, time must become space if computation is to serve as a medium. As Sybille Krämer summarizes Friedrich Kittler: ‘Wherever something is stored, a temporal process must be materialized as a spatial structure. Creating spatiality becomes the primary operation by which the two remaining functions of data processing – transporting and processing – become possible at all’ (Krämer 2006). At the same time, in computational complexity theory, space can be easily traded for time, and vice versa, see Aaronsen 2013.

[iii] In the following I will use this and other established metaphors of machine learning without scare quotes or footnotes, and thus without always making their anthropomorphising function explicit. I trust the reader to not be ‘fooled’ into thinking that these machines are human, or considered to be human by the author.

[iv] Images in neural networks are actually processed in batches for efficiency reasons. Multiple three-dimensional matrices (an image has three colour channels) are concatenated into a four-dimensional matrix which is then routed through the layers of the network.

[v] The weights in neural networks need to be initialized, but how exactly initialisation influences learning is an open question – randomisation is only one strategy among others.

[vi] This approach – which is an essential technique of contemporary machine learning – is known as learning rate decay.

[vii] In the forward pass, a prediction is made about an input image, for instance which predefined category it should be attributed to. In the backwards pass, the prediction is compared to the so-called ‘ground truth’, for instance a label containing the image’s category, and the parameters are adjusted in the ‘direction’ of the ground truth through a process called stochastic gradient descent.

[viii] Checkpoints, interestingly, usually do not include architectural information. They are representations of the state of a structure without the structure.

[ix] See Szegedy 2013, as discussed in Offert 2023.

[x] Panofsky’s (1955) distinction might seem out of place here but indeed the work required to arrive at an understanding of a foundation model is not unlike art-historical work. See also Impett and Offert 2023.

[xi] See for instance the work of scholars like Ruha Benjamin, Lilly Irani, Virginia Eubanks, Safiya Noble, or Helen Nissenbaum, to only name a few. A good introduction is provided by D’Ignazio and Klein (2020).

[xii] The ImageNet dataset, for instance, inherits its categorisation structure from WordNet, which was started with the explicit goal to produce a comprehensive ontology of what exists.

[xiii] An example popularized by a 2023 article in the New York times is a fictional 1976 version of ‘Tron’ directed by Alejandro Jodorowsky (Pavich 2023).

[xiv] ‘Die Forderung, daß Auschwitz nicht noch einmal sei, ist die allererste an Erziehung.’ Adorno 1970: 135.

[xv] In the framework of German media theory, it is of course only through technology, through

'discourse networks' [*Aufschreibesysteme*] that history can be made in the first place. See Kittler 1990.

[xvi] For a comprehensive overview see Löwy 2005.

[xvii] 'Vergangenes historisch zu artikulieren heißt nicht, es zu erkennen, 'wie es denn eigentlich gewesen ist' [...]. Es heißt, sich einer Erinnerung bemächtigen, wie sie im Augenblick einer Gefahr aufblitzt.' Benjamin 1974: 695.

[xviii] For a recent overview of the field's formation, see Raley and Rhee, 2023.

[xix] The technical term 'zero-shot image labeling' refers to the captioning of images without further training or fine-tuning a model on the dataset that contains them.

[xx] Here, I am referring to the specific, proprietary pre-trained model released by OpenAI in 2021. Since then, there have been multiple attempts to replicate CLIP in an open-source context. See, for instance, the OpenCLIP approach proposed by Cherti 2022, and research done at LAION to produce efficient pre-trained OpenCLIP models: https://laion.ai/blog/large-openclip/.

[xxi] Bruce Davidson, *Spain*, 1965. https://www.moma.org/collection/works/53558.

[xxii] Aaron Siskind, *Volcano 1*, 1980. https://www.moma.org/collection/works/45219.

[xxiii] Robert Watts, *Yamflug / 5 Post 5*, 1963. https://www.moma.org/collection/works/136552

[xxiv] I have argued elsewhere that this kind of 'humanist hacking' which resorts to metalanguage will become more common in the near future (Offert 2023b). In the meantime (early 2023), OpenAI has improved their safeguards and the hack will not work anymore.

[xxv] Jäger's images are not reproduced in this essay for ethical reasons. For a sample of his specific aesthetic facilitated by early Kodachrome film see Cosgrove (n.d.).

[xxvi] There is an argument to be made here, too, that such models, following Barthes analysis of textual contingencies, produce an estranged machinic realism. See Barthes 1982.

[xxvii] That the family is depicted as Black is a result of a superficial bias mitigation attempt by OpenAI that was exposed in 2022: random 'diversity' keywords ('black', 'female', 'asian', etc.) were added to prompts before being fed to the model, without the user's knowledge. See Offert and Phan 2022.