
VECTOR
MEDIA
IMPETT
OFFERT
DRUCKER

Vector Media

IN SEARCH OF MEDIA

Clemens Apprich, Timon Beyes, Mercedes Bunz,
and Wendy Hui Kyong Chun, Series Editors

Action at a Distance
Archives
Boundary Images
Communication
Digital Energetics
Digital Theory
Machine
Markets
Media and Management
Neural Networks
Organize
Pattern Discrimination
Really Fake
Remain
Technopharmacology
Undoing Networks
Vector Media

Vector Media

Leonardo Impett and Fabian Offert

With an Introduction by Johanna Drucker

IN SEARCH OF MEDIA

University of Minnesota Press
Minneapolis
London

meson press

In Search of Media is a joint collaboration between meson press and the University of Minnesota Press.

This open access publication was generously supported by the Canada 150 Research Chairs Program and Simon Fraser University.

Bibliographical Information of the German National Library
The German National Library lists this publication in the Deutsche Nationalbibliografie (German National Bibliography); detailed bibliographic information is available online at portal.d-nb.de.

Published in 2026 by meson press (Lüneburg, Germany) in collaboration with the University of Minnesota Press (Minneapolis, USA).

Design concept: Torsten Köchlin, Silke Krieg

ISBN (PDF): 978-3-95796-098-6
DOI: 10.14619/0986

The digital edition of this publication can be downloaded freely at: meson.press. The print edition is available from University of Minnesota Press at: www.upress.umn.edu.

meson press, Salzstr. 1, 21335 Lüneburg, Germany
info@meson.press

Contents

Series Foreword

Our present, characterized by machine-learning processes, technological infrastructures, and emerging data worlds, urges us to think about how media are shaping the conditions under which we live, socialize, communicate, organize, and learn. Searching for media as preexisting conditions of our everyday lives requires us to question the parameters, the limits, the times, the spaces of media today. How do media form to our current situation? And how does the situation make use of them?

This book series looks at these questions by examining the often hidden "terms of media" under which users operate. Rather than producing a set of explanatory keywords to describe media practices, the aim is to understand the conditions under which media are produced, and the ways in which media impact and change those conditions. Clearly, recent technical developments have led to the idea of an always-available knowledge. At the same time, this has transformed the very nature of knowledge production itself, including the assumption that more data must lead to more knowledge. Beyond the mere accumulation of data and knowledge, disciplines—from sociology to economics, from the arts to the humanities—are increasingly in search of media as a way to add context and reinvigorate their methods and objects of study.

Exploring the situation of digital media and technology, this series asks: What are the conditions of media knowledge today? To answer this question, each book features interventions by several authors whose approaches to a concept or term (e.g., archives,

communication, networks, patterns, machine, markets) diverge and converge in surprising ways. By bringing together scholars from different intellectual fields and academic regions, this series aims to advance media and technology theory by provoking new descriptions, prescriptions, and hypotheses—to rethink and reimagine what media can and must do today.

Acknowledgments

This book emerged from a series of conversations at the 2023 Schloss Dagstuhl seminar on “Visualization and the Humanities,” where we first met the brilliant Johanna Drucker, whose work had formed the basis of a recent article of ours, “There Is a Digital Art History.” We had originally planned to extend our existing collaborative work into a short book; two years later, the scope has turned out quite differently.

This book has been profoundly collaborative in more ways than one. The chapters were written entirely in parallel, and the main arguments emerged very much from the sustained dialogue between us. But we are also profoundly indebted to many other individuals for their intellectual generosity, critical insight, and support along the way. Among many others, we wish to thank: Anne Alexander, Hannes Bajohr, Peter Bell, Alan Blackwell, Mercedes Bunz, Geoff Cox, Eleanor Dare, Ranjodh Singh Dhaliwal, James Dobson, Tristan Dot, Ryan Heuser, Emmanuel Iduma, Wolf Kittler, Alan Liu, Elizabeth Mansfield, Roland Meyer, Matteo Pasquinelli, Thao Phan, Rita Raley, Eryk Salvaggio, Ludovica Schaerf, Anna Schewelew, Tyler Shoemaker, Antonio Somaini, Alessandro Trevisan, Christina Vagt, and members past and present of the Ways of Machine Seeing group and the Cambridge Critical AI Seminar.

A generous grant from the Volkswagen Foundation through the AI Forensics project, and the Max Planck Society through the Machine Visual Culture project, allowed us to work on this book in person

in Cambridge, Rome, New York State, and Santa Barbara over the course of two years. We are also grateful to the two anonymous reviewers for their careful readings and incisive comments.

Without this rich constellation of collaborators, interlocutors, and communities, *Vector Media* would not have been possible.

Introduction

Modeling Visual Dimensions

Critical Concepts and Intellectual Traditions

Johanna Drucker

What is modeled when we model vision with computational tools within automated systems? And how do we read these models as historically and ideologically inflected cultural artifacts? In *Vector Media*, Fabian Offert and Leo Impett bring their combined expertise to bear on these questions. They read the technological formations as ideological, situating their analysis among multiple fields, including media archaeology, critical theory, visual studies, and machine learning. Each concept contributes to the framework for their arguments.

First, a word about what they are not doing, which is to address the effects of machine vision from a front-end user perspective. Much good work exists in this area, focused on cultural impacts, the ubiquity of sensors, surveillance, and security systems, and the biases built into data sets as well as into the algorithms by which image recognition and classification operate (e.g., Rettberg 2023). The related field of critical data studies links these concerns to the analysis of all AI systems, not just those specific to machine vision, and emphasizes ways the materials on which automated systems are trained have built into them various problematic features that

result in discrimination against people of color, ambiguous gender identity, and marginalized positions (Buolamwini 2024; Schaaf et al. 2021). While Impett and Offert touch on issues of data bias, their focus is on the "black box" itself—the workings of the AI operations and systems in machine vision. Or, to put it differently, they are addressing the operationalization of the underlying assumptions of AI applied to visual media.

To do this, Offert and Impett draw on a critical media studies paradigm that attends to the ways models of automation are epistemological, not just technical, in their operation. They address assumptions in theories of vision and their implications for knowledge production in and through automated systems. They look at the attempts to use AI to recognize, classify, and generate visual images. They also make use of media archaeology, and its earlier precedent in what was termed "apparatus" theory in the 1970s, and the proposition that the devices, machines, and systems of media should be read as substantive artifacts rather than interpreting the specific contents of what these deliver (Cha 1980). In fact, the "delivery" or "transmission" approach to media is discounted in these approaches, replaced by careful reading of the values that are incorporated in the ways systems and devices are themselves designed and implemented. Finally, they draw on critical epistemology applied to the operations of automated systems to arrive at a theory of "post-media-specificity" and the key concept of "neural exchange value." Much is packed into this work and the critical insights that define a unique niche in the literature.

Theories of Vision

Machine vision is a deceptively simple phrase, masking the compound character of the varied intellectual and technical dimensions that are part of its implementation. Even using the concept of "seeing" is immediately problematic and requires careful pulling apart of the various strands of vision, cognition, information processing, image analysis, representation, and visual epistemology that contribute to the field. This brings us back to the first question posed

above: What exactly is being modeled when we apply computational tools to vision? The processes of perception and cognition in the visual/optical realm? Images and representations of the visual world according to pictorial or photographic conventions? Ways of making images computationally from component parts and elements? Ways of identifying and classifying visual phenomena? Visual modes of knowledge production? Each of these brings very different challenges.

What would it even mean to model "vision"? Vision is so dominant a sense in human beings that defamiliarizing it through historical and cultural filters challenges basic intuitions about the perception of the world (Lindberg 1976). Even the notion that "the world" is merely "perceived" plays into the idea of vision as fundamentally passive transmission of light through the eye to the brain (Cavanagh 2011). The influential eleventh-century scholar Avicenna said, "The eye is like a mirror, and the visible object is like the thing reflected in the mirror" (Findlen and Bense n.d.). This statement was the outcome of the accurate scientific observation that light formed an image on the retina. But it gave rise to a reductive notion that the retinal image itself traveled directly to the brain the way an image moves through a fax machine from sender to receiver. This naïve theory of vision persists despite experimental evidence to the contrary and contributions from art history and cognitive studies. Impett and Offert conceptualize their work on very different premises, but the persistence of these concepts continues in popular perception even though there is a long intellectual tradition offering substantive critique.

In one such critical approach, art historian Ernst Gombrich (1960) debunked this naïve notion of the "innocent eye" in his canonical work, *Art and Illusion*, first published more than half a century ago. He argued that vision was never simply a matter of direct apprehension of a physical world through an optical system. The act of perception is inflected by a combination of factors that contribute to what and how the visual system works. We do not simply see, which is to say, seeing is neither simple nor direct.

Most importantly for Gombrich, we see what we know to look for, and that knowledge is intimately bound to historical, cultural, and personal circumstances. Artists looking directly at the unfamiliar forms of beached whales drew features they expected to see rather than what was anatomically correct.

Gombrich's argument is far more nuanced and complex than this series of statements, of course, but it is complemented by that of another art historian, Michael Baxandall (1988), who formulated the idea of "the period eye"—a mode of seeing shaped by the era in which it occurs, much as period speech or manners betray the moment of their production through an infinite number of specific properties and details. Evidence of his observation occurs through the fact that forgeries age differently from originals. The features considered "characteristic" of an artist's work are not the same across time.

Another important contribution to a more complex theory of vision came from J. J. Gibson (1979), who drew on the earlier work of Jakob von Uexküll to create an ecological theory of vision in animal species. In view of both scientists, animals, including humans, construct a visual image of the world based on their needs, behaviors, and capacities, making use of different sections of the electromagnetic spectrum. Optical phenomena exist, but perception is a construction.

The details of visual cognition are still a matter of much debate, and whether the brain builds "models" or processes color, shape, and position separately through neuronal firings is an open subject in investigation and research. But several crucial issues are well established. One is that a degree of what is referred to as neural plasticity is part of cognitive activity, and that the elements of the anatomical apparatus—nerves, neurons, signal processing—are subject to modification through reinforcement and/or trauma. The physiology of visual cognition is not universal, nor is it independent of the experience by which we learn what to see. In this regard, models of neural networks developed to make use of feedback for

self-reinforcing and "unsupervised" learning perform similarly, but in a very different silicon-based substrate—not the living cells and systems of a body. This malleability is important because it also emphasizes the situatedness of sight within embodied historical and cultural conditions that are (still) impossible to simulate.

Picture Processing

In the design of automated systems that process visual phenomena, the physiology of vision and cognition is often set aside as irrelevant. In fact, very little actual modeling of vision is achieved with computation. Instead, as Offert and Impett make clear, "picture processing" is the chief accomplishment of automated systems performing recognition/classification and production/generation tasks. While vision is situated, embodied, and involves both peripheral and centered sight as well as the synthesis of other senses in cognitive processes, the way "machines" produce the illusion of modeling vision comes from constructing a representation that is often grounded in photographic conventions. These are generally (though not always) monocular, single-point-perspective, flat images projected on a plane that are assumed to be uninflected by experience. One has only to recall the standard illustrations of Renaissance artists, with their gridded glass situated between an artist and a model, a bowl of fruit, or a scene to recognize this trope (Panofsky 1991). Paradoxically, perspectival images are used as if they were merely mechanical and observer-independent, ignoring the situatedness of the point of view, position, distance, and scale involved.

Approaches to image recognition as an automated process have varied over the decades, like other AI developments. Some are trained to optimize feature detection, pattern recognition, and identification of other formal components through learning to match these elements. For surveillance and analytic purposes, these are methods that have efficacy—scanning for errors on an assembly line, a military target in a landscape, an anomaly in a

medical scan, and so on. Many use human beings to tag and identify the contents of images in order to train automated systems and improve performance. A picture of a "dog" was clearly in a category different from a "cat," and so distinctive features could be discerned, identified, and made to serve as the basis for recognizing other images with similar features. The long experiment known as ImageNet, launched at Stanford by Fei-Fei Li in 2006, relies on human-generated verbal tags (i.e., labeling) to establish a classification system for images. However, it turns out that distinguishing your cat Mittens from every other feline on the planet is a non-trivial task. Yet, images, not visual experience, remain the basis of analysis in these systems because their analysis can be automated.

In computational image processing, the pioneering neurophysiologist David Marr (1982) attempted to formalize what he believed were the fundamental stages of visual perception. In order to build a computational model, he mapped a progression from an initial schematic perception (shape and form) to the addition of textures, patterns, and shading, and then a full three-dimensional model. This sequence isolated individual features that could be programmed as an information system in which visual phenomena are apprehended automatically. His work is still seen as foundational for its rigorous approach to formalization of visual parameters. But Marr's vision of vision, as Impett and Offert note, lacked any insight into its historical and cultural dimensions as an epistemological system. It was mechanical rather than experiential.

Computational abilities to "recognize" images have progressed and can be automated with prompts to an increasingly nuanced and specific degree. Whether human input is used to train the neural networks, or computational analysis uses feature recognition, the mathematical models these generate for purposes of establishing similarity and difference have become increasingly sophisticated—and as Impett and Offert describe, far from visual images. Such complex mathematical analysis is what underpins what they identify as "neural exchange value" within automated systems,

a concept they describe as a "post-media" approach, one in which "image" is stripped of all materiality, replaced by a mathematical model. And yet, the inventory of visual features persists and can be used, among other things, to generate images from an inventory of component parts.

Image Generation

The idea of machine-generated images goes back at least to the development of automata capable of "drawing" or "writing" as an entertaining novelty beginning in the eighteenth century. The notion of making art computationally builds on other procedural and formal modes of production within conceptual traditions built on rule-based constraints (Drucker 2018). Alan Turing famously created a combinatoric "love letter generator" with a collaborator, Christopher Strachey, in 1952, using a Manchester Mark mainframe computer. Not surprisingly, the capacity to generate images using computer-generated programming and output was a subject of artistic and technical research in the 1960s. By 1970, three major art exhibitions featuring "computer" and "electronic" art had showcased the range of projects in this field: *Information* (curated by Kynaston McShine at MoMA in New York), *Software* (curated by Jack Burnham at the Jewish Museum in New York), and *Cybernetic Serendipity* (curated by Jasia Reichardt at the Institute of Contemporary Art in London).

These early examples, like the work that came out of Bell Labs, made use of computers in several ways, but, perhaps more importantly, demonstrated the fascination exerted by concepts like "programming" and "circuits" for aesthetic production. Some used computers graphically, as the means to translate the tonal values of images into printer icons that worked like grayscale pixels, as in the work of Kenneth Knowlton. Instruction-based procedures also quickly came into play, with pioneers like Georg Nees and Michael Noll writing algorithms for step-by-step production. Many were essentially modes of directing a plotter pen to draw lines or a dot-matrix printer to produce patterns from programmed instructions,

but information processing and code-based instruction were being linked by the 1960s (Goodchild n.d.). An important contribution came with Max Bense's theoretical work on generative aesthetics, first articulated in Stuttgart in the mid-1960s. His agenda was to resist the affective force of aesthetics he had seen at work in the rise of fascism. Machine-made art might do that (Bense 1965).

The artist Harold Cohen created a very different image-making system when he made *Aaron*, a computer-driven robotic painting machine that contained models of the visual world. The system could make use of various parameters, such as foreground and background, overlapping forms, and size relations to indicate distance. Aaron's model was unusual in its attention to visual experience, though even this was largely based on the conventions on which pictorial representation of that experience was governed.

The development of graphic programs was another step formalizing visual parameters into computationally tractable form. But as these progressed, particularly in animation and virtual reality environments, the analysis of how to create a computational language of form, surface, light, space, and even behaviors became increasingly sophisticated. Many of the animation systems were semiautomated, capable of averaging between cells to create illusions of motion or making characters and objects behave according to programmed instructions. But when the GANs (generative adversarial networks), invented by Ian Goodfellow in the early 2010s, began to generate images as hybrids, a new level of illusion appeared that has shifted into the "deep fake" domain. These systems are, again, built on neural net technologies that use reinforcement to learn. They are called adversarial because of the way they internally assess their own output and use this discriminatory process to improve its "realism." But, here again, we are back to the cultural and historical situatedness of the models, for the "real" is largely judged in terms of photographic conventions that can be best assessed in critical visual studies, as noted above.

Media Theory

Among the strains of media theory relevant to these topics, apparatus theory, which arose in the work of (mainly) French and British film theorists in the 1970s, was already focused on operations, workings, and mechanisms, rather than media contents. It also drew on the semiotic studies of cultural sign systems, structural linguistics, and psychoanalytic theories of the subject. This becomes relevant to the analysis of machine vision because of the connection to photographic conventions. The French semiotician Roland Barthes (1977, 44) famously (and mistakenly) identified photographic images as a "message without a code," a belief that remains central to the ideological operations of digital imaging. But in addition, apparatus theory engaged with analysis of the "subject" as first articulated in the psychoanalytic work of Sigmund Freud and Jacques Lacan, and their extensive analysis of the phenomenon of "interpellation" as it arose in critical theory of systems of articulation and power relations. The concept of subject positions and their construction through such elaborations as point of view, eyeline, and other formal features was crucial to apparatus theory and its uptake in feminist film theory and its critique of the male gaze. These formulations contribute to the ideological analysis of visual systems, but also to the idea of enunciation—the structuring of power relations in any discourse.

A later contribution emerged in 1980s media archaeology, which combined an adjacent strain of cultural criticism often tracked to Walter Benjamin and other figures of the German Frankfurt School at mid-century with the historical work of Michel Foucault in his *Archaeology of Knowledge* (2002). This work engages materialist media theory represented by, for one, Friedrich Kittler. The work of Jussi Parikka, Erkki Huhtamo, Wolfgang Ernst, and others emphasized the careful reading of links between political and aesthetic features of media objects and artifacts. Their work also stressed the importance of reading media systems and operations, not contents, to expose the ideological workings of devices and formats.

The point was not to ignore the substance of communication, but to expose the role of media frameworks in the production and reproduction of ideological formation.

AI and Vision

Now, we add AI systems into the mix of visual studies and media theory. As the field of AI studies has proliferated (especially since OpenAI made ChatGPT available to a broad public in November 2022), so has critical discussion of "artificial intelligence." But, as Offert and Impett point out, this discussion has a longer history. Much of the debate is trapped in the no-exit conundrum of the questions of whether machines can think, are sentient, can be creative, or will equal and/or replace human beings. The tedium of these issues is like the "Is it art if it is made by a machine?" question the authors address (and rightfully dismiss) early on in their example of the reaction generated when an AI-generated image won a prize in an art contest. Not only are these debates endlessly repetitive, but the terms on which they are formulated are not interesting, since they are rooted in a reductive binary of human/machine rather than addressing the cultural systems in which both are constructed concepts.

Offert and Impett locate their discussion within scholarly traditions of epistemological critique and media theory applied to a history of vision and representation that not only exposes the workings of images, but more profoundly, as the authors themselves state, the "logic of visual representation" as an ideological formation. This means they turn their attention to the assumptions about the links between vision and knowledge on which such systems as neural networks, the reinforcing algorithms that underpin AI operations are based. These are often left inexplicit, unstated, or rendered invisible in the technical discussions as well as the popular reception of these platforms, as if technology were independent of historical circumstances and agendas.

For instance, an important historical and conceptual break exists between GOFAI ("good old-fashioned" AI) and neural network–based approaches, with their self-reinforcing capacities. Many early AI systems were rule-based. They were premised on the assumption that this was how human behaviors worked. The linguist Noam Chomsky imagined that language was governed by such logical operations, and his position asserted that the human brain was governed by rules and their transformations. Such a system had no ability to "learn," simply to perform certain tasks that a human might—such as direct a phone caller to a particular unit in a business by sorting the responses to a query statement through a binary decision tree. A system can automate this process by determining whether someone wants "billing," "sales," or "customer service" through matching the vocabulary in a statement, and it can repeat this task ad infinitum without ever learning anything new. But machine learning involves reinforcement and feedback so that the system, for instance, might come to associate a particular voice with a preference for a particular kind of assistance, even an individual person, and "recognize" them in a call because a pattern has been established. Using statistical methods, the system might begin to predict the behaviors of an individual caller. But it would not change the rules of the system. This was an ideological, not merely intellectual, decision premised on a set of beliefs about knowledge (formal) and learning (logical) that do not conform to human behaviors, which are filled with contradictions, anomalies, and emergent behaviors.

But another important issue is that the very concepts of knowledge and intelligence—how they are understood and what the terms designate—have to be defined within the larger projects of AI and automation. If knowledge is understood as empirical—an act of perceiving the world "as it is" and being able to replicate it—then all automation will be conceived as a simple problem of the capacity to make accurate copies, do formal pattern recognition, amass huge datasets, and combine statistical processing and

machine learning systems. But if, as Impett and Offert believe, knowledge is constructed in a dynamic transaction of perception and cognition inflected by cultural circumstances, then the process itself is continually emerging, exhibiting new properties. We learn not just by adding more reference points to our mental inventory, but by rethinking and reassessing our understanding. Knowledge constructs its objects through inquiry; it does not simply grasp an a priori world.

Almost paradoxically, despite their attention to visual images, Impett and Offert arrive at the concept of "neural exchange value" in machine vision. This is described as a media-independent value defined within mathematical parameters, not visual ones. By the time they are describing the "vectors" referenced by their title, we are in the realm of abstracted, mathematically defined relationships, not images. Features, properties, and characteristics disappear in the "post-media" condition of their representation in numerical dimensions. At that point, the relation between AI and image analysis becomes a fully epistemic issue, based on assumptions about the ways the media-specific information in an image can be abstracted into a set of "vectors" or parameters.

This puts us squarely in the realm of the critical reading of values embodied in the systems under investigation. We are confronted with the insight that the way systems are designed, implemented, and, importantly, received is in accord with what we think that technology can do or be. This is the realm of the cultural imaginary, where the history of theories and machines become entwined. The ways that things are constructed and thought are codependent. As ideological phenomena, these systems become naturalized. They appear to be simply what is because ideology embeds its values as if they were inherent. The question is, in whose interests is it for something to appear natural? For patriarchy, obviously, it was the hierarchy of power in relations between men and women; in imperial colonialism, an asymmetry of authority between colonizers and colonized. In any cultural system, a dynamic of enunciation is at play, one that situates the "speaker" or active agent of the

system in relation to the "spoken" objectified by the discourse. Attending to AI and automation systems as enunciative may be crucial to keeping the structuring dynamics of power relations and operations in view. Ceding authority to "systems" is an ideological choice laden with (some unforeseen) consequences.

The insights Impett and Offert bring to this study are crucial for critical understanding of how the technical is fundamentally and always social, and how the values these systems inscribe become instrumentalized even as they disappear from view. Their contribution is a detailed analysis of the workings of automated vision systems that demonstrates how we can understand the emerging systems that abstract the world into representations according to parameters we never "see." By providing a framework to grasp these processes, Impett and Offert offer timely expertise in an area of cultural production whose implications have yet to become fully apparent.

References

Barthes, Roland. 1977. "The Photographic Message." In *Image Music Text*, translated by Stephen Heath. Fontana Press.

Baxandall, Michael. 1988. *Painting and Experience in Fifteenth-Century Italy: A Primer in the Social History of Pictorial Style.* Oxford University Press.

Bense, Max. 1965. "Projekte generativer Ästhetik." In *Computer-grafik*, edited by Max Bense and Elisabeth Walther. Edition rot. Elisabeth Walther Eigenverlag.

Buolamwini, Joy. 2024. *Unmasking AI: My Mission to Protect What Is Human in a World of Machines.* Random House.

Cavanagh, Patrick. 2011. "Visual Cognition." *Vision Research* 51, no. 13 (July): 1538–51.

Cha, Theresa Hak Kyung. 1980. *Apparatus.* Tanam Press.

Drucker, Johanna. 2018. "Amusements Électroniques." *Matlit* 6, no. 1: 27–35.

Findlen, Paula, and Rebecca Bense. n.d. "A History of the Eye." https://web.stanford.edu/class/history13/earlysciencelab/body/eyespages/eye.html.

Foucault, Michel. 2002. *The Archaeology of Knowledge.* Translated by A. M. Sheridan Smith. Routledge.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception.* Houghton Mifflin.

Goodchild, Amy. n.d. "Early Computer Art from the 1950s and 1960s." https://www.amygoodchild.com/blog/computer-art-50s-and-60s.

Gombrich, Ernst. 1960. *Art and Illusion.* Princeton University Press.

Lindberg, David. 1976. *Theories of Vision from Al-Kindi to Kepler.* University of Chicago Press.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* MIT Press.

Panofsky, Erwin. 1991. *Perspective as Symbolic Form.* Translated by Christopher Wood. Zone Books.

Rettberg, Jill Walker. 2023. *Machine Vision.* Polity Press.

Schaaf, Nina, Omar de Mitri, Hang Beom Kim, Alexander Windberger, and Marco F. Huber. 2021. "Towards Measuring Bias in Image Classification." In *Artificial Neural Networks and Machine Learning—ICANN 2021*, edited by Igor Farkaš, Paolo Masulli, Sebastian Otte, and Stefan Wermter. Springer.

[1]

Critical Technical Practice

> AI ideas have their genealogical roots in philosophical ideas. [. . .] AI research regularly encounters difficulties and impasses that derive from internal tensions in the underlying philosophical systems. [. . .] These difficulties and impasses should be embraced as particularly informative clues about the nature and consequences of the philosophical tensions that generate them. [. . .] In short, AI is philosophy underneath.
>
> —Philip E. Agre, "The Soul Gained and Lost: Artificial Intelligence as a Philosophical Project"

Are AI images art? And if so, are they good art? Can "an" AI be creative? Might we even, one day, call a robot an artist? Is AI the future or the end of the art world? These are the questions that dominate the discourse around the relationship of visual culture and artificial intelligence today. But they are tired questions; questions that have been asked, and answered, almost since the birth of computing itself (Zylinska 2020; Offert 2019).[1] Perhaps more significantly, they fall back to an aesthetic essentialism that seems philosophically deprecated in the best case, and politically fraught in the worst. "Are AI images art?"—and all its endless variations—suggests the existence of some kind of essence of art, which works of art presumably contain. This is, of course, an unhelpful fiction. It is the sphere of art, its institutional and historical setting, which establishes the work of art as such (e.g., Berger 1972; Danto 2013). One cannot remove human agency from a work without also removing the work from this institutional and historical setting. Hence the well-known successes of AI apparently conquering the art world—Obvious's *Edmond de Belamy* (sold for $432,500 at Christie's in

2018),[2] and the even less interesting Ai-Da's *AI God* (sold for $1.08M at Sotheby's in 2024)—both successful precisely because they were, transparently, marketing stunts, experiments in economics rather than computer science.

A 2022 *New York Times* headline hints at the absurdity of the current situation: "An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy" (Roose 2022). Not the Turner prize, but the Colorado State Fair's annual art competition. Many artists are indeed unhappy in this scenario, but not because the awarded picture is art, or good art; and certainly not because "the AI" showed unprecedented aesthetic agency. They are unhappy because some of their work was used, without consent, to train a machine that would produce vastly less interesting work. Such incongruities, which emerge from a mismatch of assumed philosophical significance and actual discursive redundancy, are both well-known and well-theorized (Offert 2023, 2024a, 2024b). And yet our contemporary discussion is riddled with endless resuscitations of the debate around machine creativity, stuck in a quagmire for at least a half century. If we seem unable to progress since the 1960s, then it is perhaps because we have been asking the wrong questions all along: questions that are exclusively concerned with the *output* of computing machines, the clichéd pictures they make, the mediocre poetry they write, the elevator music they compose.

The algorithm mentioned in *The New York Times* article is one of many generative text-to-image models that emerged during an explosive moment in AI research. Under the skin, many such systems are powered by a model called CLIP (Contrastive Language-Image Pre-Training), released in 2021 (Impett and Offert 2024). CLIP is not itself an image generator, but a text-image-similarity calculator. It cannot produce pictures by itself at all; instead, it is the breakthrough that enabled the "multimodal turn" in artificial intelligence. It is systems like CLIP that this book addresses: Systems that quietly work in the background, informing and determining the gamut of possible images that can be made by a machine. More generally, the book hopes to convince the reader that the

generative perspective, looking only at the *outputs* of image- or text-producing AI models, turns out to be not very "generative" at all—and is increasingly unhelpful in understanding, and critiquing, contemporary artificial intelligence. To be clear, we are not arguing here for a blanket dismissal of outputs,[3] but rather for a somewhat radical recalibration of our critical focus: from machine images to image machines, from external to internal representations, or, more precisely, from pixel space to vector space. Because it is the epistemology and ideology of vector space—as a space of universal commensurability—on which the aesthetics and politics of contemporary machine images rest. Almost all contemporary artificial intelligence systems, we argue, can be understood as *vector media*: They no longer embody a theory of vision alone, but also a theory of knowledge manifested in the central role that vector spaces play in the modeling of culture.

This book thus tries to tackle the problem of the relationship of machine vision[4] and visual culture more fundamentally. Our approach draws on *Bildwissenschaft* (often awkwardly translated as image science), a field rooted in art history, cultural theory, and media studies, which examines how images function as vectors of meaning, perception, and knowledge. We extend this tradition by adapting it to explore machine vision systems as both the producers of visual culture and the products of deeply embedded epistemological and ideological frameworks. We are interested, in particular, in the messy entanglement of the history of vision machines and the history of vision theories (Azar et al. 2021). But by following the historical mission creep of connectionist artificial intelligence—from images and words to concepts, from compression to representation to generation, from the perceptual to the epistemic—we will see how ways of thinking about images and vision have profoundly shaped all forms of contemporary intelligent media.

Just after the release of Stable Diffusion (one of the most widely used open-source text-to-image AI systems), a perplexed Reddit user asked why the system worked *even when not connected to*

the internet. The "folk theory" explanation for the capabilities of generative image models is that they scour the internet for images that fit the text, in something akin to a Google Images search, and then combine these images through a form of computational pastiche or collage. How could the downloaded model of only about two gigabytes have stored the many million possible images in its training set?[5] How could its space of visual possibilities be encoded in a file the size of a mid-length YouTube video? How, in other words—the words of the makers of Stable Diffusion—was it possible to produce "a single file that compresses the visual information of humanity into a few gigabytes"? (Stability.ai 2022). To spend less than one byte per training image?[6] The reasons behind the impressive compression efficiency of large multimodal models like Stable Diffusion are technical, certainly, but they are also deeply ideological. Revisiting Phil Agre's claim that artificial intelligence is "philosophy underneath," we argue that the philosophy of contemporary artificial intelligence no longer resides in classical thought experiments by Alan Turing or John Searle, but rather in a historical lineage of the computational compression of perception and, ultimately, knowledge; a trajectory stretching from nineteenth-century vector mathematics to 1980s computational biology, all the way to today's multimodal foundation models. Precisely as vector media, such systems start to pose philosophical, and eventually political, problems.

In the first chapter, we lay the groundwork for a critical approach to machine vision by revisiting Phil Agre's notion of "critical technical practice"—an approach attentive to the philosophical commitments embedded in artificial intelligence systems, and to the interactions between the two. Artificial intelligence, in this view, is not just a set of tools, but an epistemic framework in which concepts from philosophy, cognitive science, and media theory reappear, implicitly and explicitly, in coded form. Its technical difficulties are not the result of insufficient engineering, but symptoms of unresolved tensions just beneath the surface. Rather than evaluating artificial intelligence through its outputs, we ask what kinds of internal

structures make those outputs possible, and how they quietly reshape the terms of visual culture. The disciplines best positioned to make sense of these developments have struggled to keep pace, often constrained by inherited metaphors that fail to register the specificity of contemporary machine learning. What is needed, we propose, is an approach that treats technical architectures (Dobson 2025) as both epistemic and aesthetic objects—not to demystify them, but to read them, to parse their assumptions, and to track the conceptual work they perform.

The second chapter investigates the epistemological underpinnings of machine vision systems, framing them not merely as instruments of image production or recognition, but as models of perception in their own right. Drawing on thinkers such as David Marr and James Gibson, the chapter traces how visual knowledge is theorized within computational systems: not as passive registration, but as a process of inference, construction, and selection. Machine vision does not simply simulate human sight, but enacts a different perceptual logic altogether—one built around filtering, optimization, and compression, although these are themselves often neurologically inspired. Indeed, the titular claim of machine vision—that it is a model of vision—is misleading. Vision models, more precisely, contain an immanent theory of *pictures* (Impett 2024). This perspective, the chapter argues, calls for an expanded image theory. If traditional *Bildwissenschaft* studied the image as a cultural form, what is needed now is an account of the image machine: a critical framework for understanding how vision itself becomes a modeling problem, governed by logics irreducible to either human perception or computational logic exclusively.

At the heart of contemporary machine vision, however, lies a machine learning technique called *embedding*, which we examine in depth in the third chapter. Embeddings (the epistemically compressed data that result from the process of embedding) are now conceived as a tool not only for representing knowledge, but for producing it—a development some computer scientists have in fact framed as a post-structuralist turn. Embedding enables

the creation of vector spaces that establish relationships between cultural artifacts. But to establish such relationships, these cultural artifacts first have to be made commensurable. This technical-ideological shift also underpins the recent evolution of multimodal foundation models, which appear to transcend media boundaries by encoding text, images, and even audio within a single model. If machine vision was implicitly modeling specific media (especially photography) for much of the twentieth and twenty-first century, it is now modeling a—still specific but highly ideological—media indifference.

This *media collapse*, theorized in chapter 4, signals a new mode of valuation: a form of value that cultural objects acquire once they enter a multimodal embedding space, where media distinctions dissolve into a commensurable—and therefore commercially exchangeable—space. We thus propose to understand "foundation models" (which are intended to be able to perform well across *any* visual task) and "multimodal models" (which can take texts, images, and often other data modalities as inputs or outputs) as facilitating a universal equivalence of cultural objects, which we, by way of analogy with the terms of political economy, call *neural exchange value.*

In that sense, the book moves from a broadly phenomenological theory of vision to a political theory of media—and, in doing so, mirrors the historical arc of its object.

Breaking the Feedback Loop of Critique

At this point we will return, for a moment, to the unhappy artists from *The New York Times* article quoted above, to better understand the feedback loop of critique we are aiming to break up with this book. The idea of "fooling" a human jury serves as a kind of litmus test for artificially intelligent behavior. Computational mimesis: Our test for good intelligence is the same as our test for good art (or at least what it used to be), the artificial passing as the natural, as in the mythical competition for mimesis between Zeuxis and

Parrhasius.[7] A litmus test that, of course, can be traced back not only to 1960s computer art, but to Alan Turing's famous 1950 paper, in which he proposes the Imitation Game—now better known as the Turing Test[8]—to tackle the question of whether machines can think.[9] Computer scientist Scott Aaronson (2013, 34) has stated that 70 percent of "everything that's been said about artificial intelligence" is somewhere in Turing's paper, and the question of whether a machine can be creative is no exception.

The question of machine creativity, of course, has been asked well before Turing. Ada Lovelace considered it already in 1843, in the paper in which she also composed the first computer algorithm. There, she suggests that Charles Babbage's Analytical Engine, arguably the first computer even if it was never built, might be programmed to "compose elaborate and scientific pieces of music" (Menabrea 1842). But Lovelace was clear that the Analytical Engine had no "pretensions to originate anything." In a section titled "Lady Lovelace's Objection," Turing (1950, 450) thus argues that Lovelace would only have been thinking of predictable, rules-based algorithms, of the sort she herself had invented.[10] A "child machine," an artificially intelligent machine that could *teach itself* (454) would, for Turing, be quite a different matter. For him, describing the anthropomorphic behavior of such a machine as "intelligent" (or even attributing "thought" to it) is the consequence of simply affording subjectivity to other humans, a natural extension of the "polite convention that everyone thinks" (446).[11] In other words, as long as our measure of intelligence is based on simply talking to people, there is no reason why we would not ascribe intelligence to machines that talk to us also. Turing's solution to Lovelace's (imagined) objection mirrors his general methodological pragmatism: A sufficient operationalization of a human faculty should be regarded as a sufficient definition of that faculty.

And it is this methodological pragmatism that would continue to shape the evaluation of generated images much later. Humans cannot distinguish whether a human or a machine made this picture or wrote this poem[12]—so goes today's version of Turing's

argument—as it has fooled a colleague dragged from the hallway, an undergraduate student, a small town art competition jury, a random sample of internet strangers, and all of them even *preferred* it over human-made works of art. This default position toward cultural artifacts generated by artificial intelligence systems, one in which both empirical "proof" and hermeneutical critique starts from the output of a system, stands almost exclusively in the shadow of Turing, to the point where we could describe any such endeavor itself as a trivial imitation of the Imitation Game. What is more, contemporary commentators (with only a few exceptions, e.g., Gaboury 2013) have been quick to forget the gender politics of Turing's original formulation. The game Turing describes differs slightly from the simplified version we tend to remember: In it, a computer, a man, and a woman are all trying to convince the questioner that they are a woman, and the computer's challenge is to fool us at least as well as the man. Elisabeth Mansfield (2007) reminds us that gendered ideologies play a central role in the history of mimesis: Tasked with making a portrait of Helen of Troy, Zeuxis is said (by Cicero) to have combined the most beautiful features of five women of Croton. We can understand the functioning of artificial intelligence systems in this historical tradition of representation, depiction, and mimesis, as we shall see in subsequent chapters.

If Turing's thought experiment provides the theoretical and practical framework to all but a few attempts to operationalize creativity, the *critique* of this Turingist position is informed, again almost exclusively, by another well-known philosophical thought experiment: philosopher of mind John Searle's Chinese Room (1980),[13] which also evaluates a system by fooling an external observer, but this time the deceit acts as a proof against mechanical intelligence. The thought experiment, part of a wider philosophical argument against computationalism—the assumption that mental states can be sufficiently modeled by computational states—posits that "intelligent" behavior might appear from a system that cannot actually process "meaning." Searle, an English speaker who doesn't

understand non-European languages, imagines himself in a room with a large dictionary of "Chinese" phrases and the appropriate responses. By following instructions in English, he can look up and produce appropriate responses in "Chinese," fooling those outside into believing he understands the language. Consequently, the Turing Test cannot be a sufficient measure of intelligence. What Searle proposes instead as a measure of intelligence is intentionality, a philosophical concept first proposed by Franz Brentano and later famously taken up by his student Edmund Husserl, which describes the directedness of a thought toward something other than itself, be it material or immaterial, internal or external. And much like Turing's Imitation Game, the Chinese Room lives on.

One of the most recent, and prominently received, variations of it appears in Emily Bender and Alexander Koller's paper "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." Bender, a computational linguist who went on to write the influential "Stochastic Parrots" paper, which offers a general critique of the usefulness of large language models, argues on a somewhat more philosophical level that the task of training large language models, "because it only uses form as training data, cannot in principle lead to learning of meaning" (Bender and Koller 2020, 5185). To illustrate this argument, the following thought experiment is introduced: A "hyper-intelligent deep-sea octopus" intercepts the conversation between two island dwellers A and B and secretly enters into the dialogue, answering instead of B. Given this, Bender and Koller ask, "Can [the octopus] successfully pose as B without making A suspicious" by learning "statistical patterns" in the conversation over time? The answer they give in the paper is yes, as long as it is not required for the conversation to be "grounded in the particulars of the interlocutors' actual physical situation" (5188). What else is this but a version of the phenomenological intentionality argument: the argument that meaning is deeply connected to the human faculty to relate to things in a meaningful way, to operate with "aboutness," a connection to the real world that goes beyond symbolic representation.

Searle's thought experiment itself, unsurprisingly, stands in a long philosophical tradition. Soviet cyberneticist and writer Anatoly Dneprov (1961) developed a formulation very similar to Searle's, in which the participants of the Soviet Congress of Young Mathematicians collectively emulate (using only zeros and ones) a machine translation algorithm between Russian and Portuguese, a language none of them knew. Gottfried Wilhelm Leibniz had already proposed a thought experiment in section 17 of the *Monadology* from 1714, in which a mechanical mill acts as a brain.[14] His thought experiment, in turn, can be traced back to the original Cartesian problem of the location of uniquely human faculties—which we can find, closing the historical loop, almost verbatim in Turing. In his refutation of the "argument from consciousness," Turing (1950, 446) quotes Geoffrey Jefferson (1949, 1100), who claims that it is not sufficient for a machine to write a sonnet to be considered intelligent, but that it would also need to be able to relate internally to its writing, that it would need to "know that it has written it," and is not simply acting "in parrot fashion." Jefferson (1949, 1106) makes explicit that this expression is, of course, a reference to René Descartes's (2006, 47) original argument that "magpies and parrots can utter words as we do, and yet cannot speak like us."[15]

Turing's response to this challenge is interesting—and will, perhaps surprisingly, catapult us right back into the present. Because Turing essentially argues that a dialogical form of examination would likely convince Jefferson that the sonnet is *understood* rather than just repeated "in parrot fashion." This *viva voce* (in Turing's terms) is, by no coincidence at all, exactly what large language models are "asked" to do. At the interface level on the one hand, but also, and at the same time, further down the stack, in an approach now widely known as chain-of-thought prompting (Wei et al. 2023) meant to enhance and evidence their reasoning capabilities.[16] Searle, of course, would have none of this—after all, to him, there would be no difference between one-off answers and entire chains of thought simply because both would emerge from the same "dictionary," the same book of transition rules.

We are thus facing here—as already in Turing and Searle—a peculiar *feedback loop of critique*, where we can infinitely expand the requirements—both quantitative and qualitative—according to which certain outputs would be understood as evidence of "thinking" (or its diffusely weaker form, reasoning). This "thinking" is taking place, as long as we also infinitely expand the dictionary from which these outputs are generated. For the neo-Turingists, then, the emphasis remains strictly on observable behavior—on mimesis, the machine's ability to mimic humanlike responses and deceive a human observer. Since only surface-level behavior is deemed relevant, the internal workings of the system are treated as neglectable to its "intelligence." In the Searlist critique, on the other hand, the very notion of meaning is inherently human: by definition, something a machine cannot produce or possess, no matter what internal processes it performs, or, more precisely, how extensive its dictionary is imagined to be. The Turingist and Searlist perspectives, then, for different reasons, end up in the same place (which, in a sense, is still the place of Descartes and Leibniz): that examining a machine's internal structure is futile or irrelevant, sidestepping any meaningful examination of the internal logics of artificial intelligence systems, and the ideologies that are—we suggest—quite literally embedded in them.

How do we break this feedback loop of critique? We would like to suggest that one way is to be quite a bit stricter—that is, think more concretely about the objects of Turing's and Searle's respective thought experiments. And indeed—a fact we need to be constantly reminded of—for both Turing, in 1950, and Searle, in 1980, the problem at hand is very much conceptual. Contemporary artificial intelligence systems operate on a basis that Searle could hardly have imagined. His metaphor for the internal operation of such a system—looking up prepared responses in a large book of Chinese phrases—might be seen to have useful analogies to the rules-based artificial intelligence systems of the last century. But not only are neural network–based language models rather different in their internal logics, the logocentric computer programs

imagined by Turing and Searle alike are now swiftly becoming multimodal vision-language systems. In Searle's thought experiment, the man in the Chinese Room has no way to connect the symbols he manipulates to any sensory experience.[17] Contemporary multimodal systems can almost imagine (so to speak) what a word *means* by relating it to pixel values: a capacity demonstrated in, for instance, generating an image from a text prompt. We have a kind of perverse embodiment, in which language tokens are grounded in plausible light-sensor values.

Mediality matters after all. It matters, in other words, how exactly Searle's dictionary is written, how it is organized, and how it is read; how knowledge enters the system (how a neural network is trained), how knowledge is stored in the system, and how knowledge can be retrieved from the system.[18] And it matters how that knowledge is originally represented—in image form, in text form, in sound form, and so on—how it is represented once it has entered the system, and how it is represented when it is presented to us as a result or solution to a task given to the system. And if mediality matters, then implementation matters. It makes a difference where exactly "the model met the machine," as historian of science Stephanie Dick (2015, 630) has argued. Searle (1980, 424) must have already been aware that this problem was on the horizon when he wrote, "'Could a machine think?' My own view is that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. And that is the main reason strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines." The core of the machine consists of a mechanism that associates inputs with outputs, both Turing and Searle agree—but Searle is thinking about this mechanism in terms of a dictionary, whereas Turing addresses it in terms of a Turing machine. The Turing machine, in turn, already contains the human element, as it is understood as an operationalization of a human computer, a person performing basic protomathematical

(indeed, set-theoretical) operations (Turing 1936). The exact nature of that mechanism, however, remains entirely immaterial in both formulations. And it is because of this immateriality of Turing's and Searle's shared reference object on which their respective positions rest, that contemporary critiques of artificial intelligence systems remain trapped in the same feedback loop of critique. Bender and Koller's octopus is meant to critique a *specific class* of large language models (not even one specific model), and yet it inevitably regresses to a critique of modeling *in general.*

However, while the *critique* of machine vision seems stuck in a feedback loop, the past decade or so has seen a radical paradigm shift in the *practice* of machine vision. Almost an entire arsenal of computer vision algorithms, painstakingly handcrafted by generations of researchers, has been replaced (practically overnight, in research terms) by one universal algorithmic paradigm: the deep neural network. This has crucial consequences not because of the "new" images generated by such networks, but because of the indeed new internal logics that allow such generations in the first place; not only because of machine-generated visual culture, but because of the processing,[19] modeling, and compression of existing visual culture; not only because of machine-made images, but because of image-making machines. Our inquiry, then, is directed explicitly at the politics embedded within the internal logics of these machines. What ideologies, assumptions, and power structures govern the processes and representations that underpin them?

Philip E. Agre's Critical Technical Practice

The eternal recurrence to Turing and Searle—and Turing's and Searle's recurrence to Descartes—demonstrates what Phil Agre (2005, 155) had lucidly stated about artificial intelligence already in the late 1990s: "AI is philosophy underneath." This statement seems intuitively true—which is why the question of machine creativity from the 1960s continues to haunt us and why it is, time and

again, answered with a version of the Turing Test or the Chinese Room. But Agre was pointing to the philosophical scaffolding of AI in a more specific way, alluding not to vaguely philosophical sentiments or thought experiments, but to actual theories that could be almost forensically uncovered through a close analysis of the concrete technical roadblocks in the technologies they underpin. For Agre, it is precisely in the cracks, the glitches, the slips of systems that we can point to their underlying ideologies. Importantly, however, this is not a one-off effort, but a systematic endeavor that must take the shape of a kind of grammar of error, to speak with Matteo Pasquinelli (2019). "AI research," Agre (2005, 155) writes, "regularly encounters difficulties and impasses that derive from internal tensions in the underlying philosophical systems. These difficulties and impasses should be embraced as particularly informative clues about the nature and consequences of the philosophical tensions that generate them."

We would like to take Agre's proposal—which was largely targeted at the previous generation of "good old-fashioned" artificial intelligence systems—to bear on the world of contemporary machine vision.[20] Indeed, we would like to propose that in recent decades[21] machine vision and visual culture have become increasingly more entangled, to the point that we now have to speak of and think about a visual culture that unfolds exclusively *within* our machines: a *machine visual culture.* If, as Agre suggests, AI embodies underlying philosophical assumptions, and yet we contend that these assumptions are not merely derivatives of Turing and Searle's thought experiments, then what theoretical, and, by extension, critical frameworks, truly underpin this machine visual culture?

Our key intervention in answering this question is to turn to a history of the technical disciplines involved in the development of contemporary machine vision, and in particular to take a second look at artificial intelligence and machine vision research during the "AI winters": the decades between the decline of cybernetics-inspired paradigms (often tied quite precisely to Marvin Minsky and Seymour Papert's 1969 *Perceptrons* book),[22] and the beginning

of the contemporary focus on deep neural networks, usually associated with the DeepVision entry to the 2012 ImageNet image classification challenge (Krizhevsky et al. 2012). There is a curious historical paradox here, in which many of the key technical innovations in deep learning and computer vision (e.g., convolution, backpropagation)—and many of the tacit assumptions about vision and knowledge that underpin them—were first proposed in the last three decades of the twentieth century, a period largely overlooked by critical AI studies, if not by the technical researchers themselves. Our task, then, is to trace how the epistemic paradigms at work in these fields form a longer arc, especially considering what we might call the "long 1980s" of artificial intelligence research.

Agre received his PhD in computer science from the MIT Artificial Intelligence Laboratory in 1989, and went on to teach at Chicago, Sussex, and finally UCLA, where he worked until 2009. During that time, Agre produced what we propose to understand as perhaps the first critical theory of artificial intelligence. To quote Michael Travers (2009, quoted in Masís 2014), Agre was "a leader of the situated action insurgency in AI, a movement that questioned the foundations of the stale theories of mind that were implicit in the computational approach. Phil had the ability to take fields of learning that were largely foreign to the culture of MIT (continental philosophy, sociology, anthropology) and translate them into terms that made sense to a nerd like me." Agre's only major book, *Computation and Human Experience*, is both a technical investigation of state-of-the-art machine learning techniques at the end of the 1990s, and a highly philosophical investigation of the inherent limitations of these techniques from the perspective of a theoretical framework—in which Agre includes elements from Heideggerian existentialism via the work of Hubert Dreyfus (2007), but also fragments of Frankfurt School critical theory and poststructuralist analysis. Yet Agre's is not solely "critical" in the sense of being descriptive, historical, or theoretical. The book is also an engineering treatise, an attempt to build a different form of artificial intelligence. Mirroring Agre's methodological concern

with a unity of theory and practice, it is both a manifesto for his proposed methodology for research into artificial intelligence—a *critical technical practice* (Agre 1997)—and a central example of it.

One of the main goals of bringing together these different philosophical discourses, for Agre, is to expose exactly those Cartesian roots of AI that are also latent in Turing and Searle. In an early essay (Agre 2005), Agre argues that from the first AI experiments in the 1950s to the expert systems of the 1990s, AI has been dominated by Cartesian dualism. He contends that the field of AI had sought to "domesticate the Cartesian soul into a technical order in which it does not belong" (Agre 2005, 13). While Cartesian reasoning itself can be mechanized, Agre argues, its role as the "soul" in cognitive architecture is fundamentally flawed and unsustainable. Agre's critique of rules-based artificial intelligence was, of course, a product of its time. Hence, we must separate the approach from today's technical specifics. To show that Agre's critique is as timely as ever, we ask: What (if not Cartesianism) are the underlying philosophical systems that underpin today's machine vision systems?

For Agre, long-standing internal contradictions of philosophical systems appear in the moment of their operationalization, in the moment in which there is an attempt to transform them into computer programs running on machines in the real world. Concrete, material, technical approaches to AI fail because of abstract, immaterial, philosophical impasses. It is from this insight into the deep entanglement of the conceptual and the technical that the notion of critical technical practice emerges in Agre's writing. "[The] point [. . .] is," as he writes, "to expand technical practice in such a way that the relevance of philosophical critique becomes evident as a technical matter" (Agre 1997). With this, Agre provided the underpinning for what was later called Nouvelle AI, pioneered by the Australian roboticist Rodney Brooks, with whom Agre wrote his PhD dissertation in the late 1980s. Brooks saw that computers that could perform highly specialized operations at superhuman level (such as accounting) were woefully inadequate at basic animal tasks (such as navigating a room). In a seminal paper from 1990

called "Elephants Don't Play Chess" (Brooks 1990), he claimed that AI research had been wrong to reduce all intelligence to the manipulation of symbolic information. Instead, Brooks and his group sought to create machine intelligence that was grounded in its physical context—many of their early prototypes resembled electronic insects. Out of his work in embodied AI, Brooks went on to invent the Roomba vacuuming robot. Yes, that Roomba vacuum robot.

"Elephants Don't Play Chess" cites one of Agre's earliest works as an example of Nouvelle AI: *Pengi*, a computer program that plays an arcade video game (*Pengo*, very similar to *Pac-Man*) using what at the time was a completely new approach to spatial reasoning (Agre and Chapman 1987). Agre, there, differentiates between the "capital-p planning" of traditional AI (wherein a smart "planning" phase constructs a plan, to be carried out in a mechanical fashion by a dumb "executive" phase); and "lower-case-p planning," a kind of "continual improvisation," informed by prototypes, recipes, and directions, but also based on how the world is now. *Pengi*, which was also the basis of Agre's PhD dissertation and a central example in *Computation and Human Experience*, is an algorithmic implementation of a theory of activity and planning as an inherently distributed, bottom-up, emergent form of computation, rather than something that can be achieved centrally. *Pengi* is an implementation of this theory as a computer program, but its point is actually not to play the game *Pengo* as well as possible. In fact, Agre admits it does a fairly bad job of that. Agre's "AI is philosophy underneath," then, goes both ways: *Pengi*'s role is to act as a kind of computational thought experiment that operationalizes a specific approach to activity and planning, with the ultimate goal of modeling a theory of how humans navigate spaces.

Agre (1985) traces his radical epistemology of AI back to two key sources: Jean-Paul Sartre and David Marr,[23] combining Sartre's observation that the workings of the mind are beyond our direct conscious access with Marr's computational theory of vision. As we will explore in the subsequent chapter, Marr's framework—

a cornerstone in the neuroscience of vision—has positioned him as a foundational theorist for contemporary computer vision. Agre's approach, in other words, is built on a complex network of (philosophical) discourses and (technical) methods; and on keeping track of how the two intertwine. In attempting to trace the epistemologies and ideologies of machine vision, we are dealing, then, not with a linear progression of theories or technologies, but with a deeply intertwined history of both. These cross-contaminations and abrupt reconfigurations—where a theory reshapes a technology and a technology recasts a theory—uncover the dynamic undercurrents of machine vision's development, and ultimately a significant paradigm shift from a theoretical and technical perspective that understands the intersection of machine vision and visual culture as a problem of *vision*, to a theoretical and technical perspective that understands it as a problem of *knowledge.*

Dataset Critique Is Not Enough

Where, then, might we find traces of *machine visual culture*, and how might we go about analyzing its epistemologies and ideologies? A logical starting point is to inquire about the nature of the images that a machine vision system has been exposed to—the dataset it has been trained on—and the information it has assimilated from them. What can easily be found, in almost all established datasets, is, in the words of Abeba Birhane (2021), "misogyny, pornography, and malignant stereotypes," and everything in between, on top of a truly intractable network of data privacy and copyright violations. Consequently, "dataset critique," over the past decade or so, has developed into the dominant methodology informing critical readings of machine vision systems. This perspective culminates in positions that view the trained model primarily as a reflection of its dataset—an "infographic" of everything the dataset contains, while inevitably constrained by what it excludes (Salvaggio 2022). "More than any other thing, [the dataset] will influence what your model can do and how well it does it," as

Christo Buschek and Jer Thorp (n.d.) write in a widely lauded piece for the Knowing Machines project.

While we do not discount such positions, we would like to argue that the methodological dominance of dataset critique has obscured the fact that datasets are not the only locus of ideology. 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 "clickworkers" in the Global South (e.g., Irani 2015), the energy and water consumption of data centers, the entanglement of artificial intelligence corporations with neofascist politics (Katz 2020; Golumbia 2024), the privacy and copyright concerns connected to indiscriminate data scraping practices, or the extractivist nature of model production in general. Instead, our analysis, and our critique of dataset critique as a dominant practice in the humanities, focuses on the model itself.

In that sense, dataset critique could be regarded as another consequence of the dominance of Turingist critique: If it is not the output that "contains" the ideology, it surely must be the input. But again the process—the pipeline from input to output—is ignored. As computer scientist Sara Hooker (2021, 1) 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. With the advent of foundation models and corresponding datasets, this becomes especially relevant: 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. 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. In fact, the most significant recent general critique of a LAION dataset (Birhane 2021) uses exactly this CLIP-powered tool to identify the

"misogyny, pornography, and malignant stereotypes" on which its main argument rests.

While datasets that move well beyond a billion images, like LAION, are now the baseline of machine vision research, for the majority of the past decade, the primary dataset utilized for training large neural networks has been ImageNet, making it the single greatest target of critical attention to training datasets. Comprising approximately fourteen million images, ImageNet has served as a foundational resource in the development of computer vision models. Fei-Fei Li, the principal architect of ImageNet from Stanford University, described the project's ambition as mapping the "entire world of objects" (Gershgorn 2022).

Foundational datasets like ImageNet influence machine visual culture directly through the use of image classification models trained on it. But they also exert indirect influence in a number of ways. In the case of ImageNet, for instance, through "transfer learning," where models trained on ImageNet are then fine-tuned on more specialized tasks (whether factory automation or plant identification); through architectural innovations in neural network design intended to "solve" ImageNet specifically due to its privileged status in computer vision research, which has outlasted the dataset itself; and through a host of ImageNet-trained subnetworks used in the training of subsequent networks, not through fine-tuning but as metrics (e.g., the perceptual distance score and Frechet inception distance) for assessing the performance of computer vision tasks which no longer use ImageNet as a basis. ImageNet, then, is still everywhere—but not as an easily browsed, and easily critiqued, set of stock photography, but as an integral part of large and complex contemporary machine learning pipelines.

As Agre understood, AI's philosophical systems are best glimpsed through the cracks and glitches in AI systems. In 2019, a Facebook study on the Dollar Street dataset, which contains images of common objects across diverse global contexts, found that their object detection model (trained on ImageNet) accurately identified items

like soap or toothpaste well in the Global North, but misidentified similar objects far more frequently in the Global South, especially in cases where Western-style domestic scenes (e.g., tiled bathrooms) were not present (De Vries et al. 2019). That same year, another paper found that machine vision systems trained on ImageNet were beginning to fail to recognize certain classes because "classes such as car or cell_phone have changed significantly over the past decade" (Recht 2019, 4). The technical solution to these "road-blocks" (to speak, again, with Agre) is relatively uncomplicated: We should diversify datasets, so the argument goes, by adding more photographs of household objects from the Global South, or by reupdating consumer goods as their visual form changes, in an endless quest to remap the "entire world of objects." If this correction proves too costly, it might be simulated automatically (for instance, by synthesizing the underrepresented images computationally); and if even that proves difficult, the biases one needs to be aware of might be captured on a so-called model card (Gebru et al. 2021). Yet this does nothing to deconstruct the fundamentally Fordist image theory of ImageNet, in which images are made up of industrially manufactured commodities to be classified. Among the one thousand categories of ImageNet[24] we find "Polaroid camera," "iPod," and "Model T" (Impett 2024).

Conventional computer-scientific approaches to dataset bias and critique, then, quickly find themselves in a fundamentally extractivist position—in the words of Johanna Drucker (2011), from *data* to *capta*—which mirrors the colonial mineral and labor extractivism so fundamental to artificial intelligence in the first place. When this argument is extended to the question of bias in face recognition (one of the key topics for technical research on bias, given the racialization of surveillance, particularly in the United States and Israel), we face not only the problematics of data capture but of the desirability of removing bias in the datasets, and therefore in the finished computer vision systems themselves. But who benefits from an increased capacity for the automatic surveillance of nonwhite faces? (Benjamin 2019; Chun 2024).

Artists and the Dataset

A less solutionist line of dataset critique has come, instead, from artists. The year 2019 saw a host of artworks tackle ImageNet: Trevor Paglen and Kate Crawford's *ImageNet Roulette* (Fondazione Prada; see also Crawford and Paglen 2019), which focused on the dataset's problematic ontologies; and Trevor Paglen's *From Apple to Anomaly* (Barbican) and Nicholas Malevé's *Exhibiting ImageNet* (Photographer's Gallery; see also Malevé 2020), *both* of which attempted to exhibit in a gallery setting a large subsection (in Paglen's case spatially, in Malevé's temporally) of the dataset. A few years later, the Brazilian artist-scholars Gabriel Pereira and Bruno Moreschi extended this to a form of mail art, sending two or three images from training datasets to a number of other artists and scholars, whom they would interview after a few weeks of "slow looking" (Pereira and Moreschi 2023).

There is certainly much to be gained from dataset critique, both theoretically and pragmatically—and we think there is particularly fruitful ground in considering the hidden labor of image datasets and in the ideologies of the labeling categories themselves. But the conventional view of bias in computer vision traces the visuality of the machine—all its culturally, socially, politically situated ways of seeing—to the dataset. This perspective is captured by a popular maxim in computer science: *garbage in, garbage out.* We might read this as an extension of our notions of the cultural situatedness of human visual and artistic perception. Michael Baxandall's notion of the *period eye*, perhaps the most famous such theory in the English-speaking world, was outlined in a book titled *Painting and Experience* (Baxandall 1972). Baxandall's theory explicitly draws a distinction between nature and nurture: between our common human neural wiring, and the specific visual cultures to which we have been exposed ("experience"). That a machine vision system should effectively inherit a period eye from its training dataset is the natural extension of this view.

Yet this is only part of the story. Indeed, this already-uncomfortable distinction between nature and nurture completely collapses in the case of machine vision, where the wiring is itself a socially and historically embedded technology. This "wiring" is a central protagonist of our work—but it is in fact plural, a large family of different neural architectures, each developed to solve particular problems. Few of them are meant to accurately simulate what we understand to be happening in the human brain with regard to visual processing. Some of them are, however, inspired by what we know, or thought we used to know, about biological vision, as we shall see in the subsequent chapter. Just as training datasets—assemblages of images, labels, correlations—collectively form a way of seeing the world, neural architectures also embody theories about how images function; they also contain a way of seeing.

What is more, dataset critique concerns itself almost exclusively with *labeled* data, where images are explicitly categorized and classified, as in ImageNet, building on an epistemology of classification (Bowker and Star 1999). Yet many computer vision systems bypass labels altogether in what is called "self-supervised" learning, where a system is trained on raw images, without predefined categories—for instance, by rotating the image and asking the neural network to discern the "correct" orientation, or by rearranging images as puzzles for the model to solve. Current practices of dataset critique have little to say about networks trained in this way, and yet they inarguably encode assumptions about visual composition, perhaps even implicit models of media. A network might learn the "correct" orientation of a smartphone photograph (with the ground conventionally at the bottom), but the task would be nonsensical for microscope slides or satellite images. Furthermore, so-called data augmentation layers—geometric transformations like flipping, cropping, zooming, and warping—impose an implicit understanding of perspective, symmetry, and other visual forms.

Architectural and algorithmic assumptions, then, and not just datasets, encode implicit theories about how images work, shaping

how machine vision sees alongside their training datasets. We can take this thought to the extreme by considering a relatively recent system trained by a Japanese team entirely without the use of so-called natural images—that is, without photographs, drawings, renderings, or any other visual training datasets (Kataoka et al. 2020). Instead, the system was trained on fractals, simple mathematical functions that result in highly complex patterns when visualized in two dimensions, which have often been compared to natural forms (such as ferns, coastlines, or broccoli). The system, thus, is trained not only without labels, but without data in the conventional sense, yet it still manages to learn image structures so effectively as to be competitive with networks trained on ImageNet for fine-tuning. We have, here, the limit case of our thought experiment: a machine vision system that has never seen a photograph or a label before—never *seen*, period—and yet manages to "solve" the task of image classification. Can we still speak of a visual culture of machine vision in this case? And if not in the dataset alone, where else—in terms of both discipline and methodology—might we go to find it, to unpick it, to study it?

We are, then, still interested in datasets, but not merely as discrete and direct packets of cultural situatedness. Datasets carry a long historical shadow, often framing research questions and technical paradigms that extend far beyond their immediate use. Bias, therefore, goes beyond representational content; it is embedded in the logic of representation itself. To fully understand the visual culture within machine vision, we need to move past training data to examine the black boxes themselves and the epistemologies encoded within their operational frameworks. This also means taking seriously the philosophically overloaded technical terms of computer science: representation, compression, knowledge, and so forth.

From Machine Images to Image Machines

Ours is certainly not the first attempt to try to cross the field of machine vision with that of visual culture. We have already touched

on the emergent practices of dataset critique; but many academic disciplines have looked at different sites of intersection between the two, and we will attempt to draw on many of them. However, we must first distinguish the kinds of questions that have been asked—all of which are relevant to the problem of machine visual culture, but none of which have quite tackled it head-on. In this chapter, we want to gesture at some of these questions explicitly here, and will ask the reader to return with us to the disciplinary question in the subsequent chapters.

Art historians and visual studies scholars have begun to look extensively at those recent additions to visual culture that we are bracketing in this book, emphasizing the aesthetic rather than the philosophical aspect of the question of machine creativity. The infamous *Portrait of Edmond Belamy*, created by a generative adversarial network, often serves as a starting point to identify the visual limitations of machine-generated works (e.g., the "GANism" style popular in the mid-2010s), and to critique their necessary dependence on the—Western, digitized—canon, often through the lens of established categories such as kitsch or nostalgia (Meyer 2024; Steyerl 2023). Related to these approaches is the exploration of the latent space offered by generative models as a space of discovery or creativity (Somaini 2025; Offert 2024a). Scholars of modern and contemporary art in particular are interested in the question of postphotography, and how concepts from the theory of photography—often building onto the work of Benjamin (1999), Flusser (2013), Barthes (1981), or Virilio (1994)—might be applied to generated images (McQuire et al. 2024; Zylinska 2023), or how such images might fit into established art historical categories (Somaini 2023). Finally, rather than treating digital aesthetics as isolated phenomena, scholars like Meredith Hoy (2017) or Shane Denson (2023) investigate aesthetic and political continuities that destabilize the analog-digital divide (see also Offert 2025a, 2025b).

Media studies—in both its North American and European incarnations—has established a variety of critical approaches for questioning the ontological status of digital images. The concept

of the operational image, introduced by Harun Farocki and later expanded by Jussi Parikka (2023) and many others (e.g., MacKenzie and Munster 2019; Offert and Phan 2024), opens a framework for understanding images not as representations, but as active components within automated systems—an approach that redirects attention to the functional role of images in technological operations and control. Jacob Gaboury (2021) and James E. Dobson (2023), meanwhile, adopt historical and genealogical perspectives, investigating how the technical architectures of computer graphics and computer vision, respectively, embed ideological frameworks inherited from their origins in interactivity and militarization. Yet, while these approaches reveal critical intersections between machine vision and visual culture, they leave the epistemic structures and operational logic of *contemporary* machine vision itself largely unexplored.

The emerging field of critical artificial intelligence studies (Raley and Rhee 2023), closely intertwined with science and technology studies (STS), and its distant but more technical cousin, FAccT (fairness, accountability, and transparency), reflect distinct sets of interests concerning machine vision and its social implications. STS and adjacent scholars have focused broadly on the societal implications of both consumer-facing and military-industrial machine vision, such as in the form of machine vision systems in Google search (Noble 2018), surveillance systems (Zuboff 2019), facial recognition systems (Meyer 2021, 2024), and militarized drone vision (Richardson 2024). FAccT scholars, on the other hand, commonly engage with AI systems at an algorithmic level. Unfortunately, however, their perspective often remains focused on issues like categorical bias or system transparency, and the constraints of scientific discourse largely prevent them from fully discussing (or perhaps, in some cases, even acknowledging) the broader cultural and ideological frameworks in which these biases operate. Research in mechanistic interpretability—the even more technical version of FAccT—often tries to identify particular visual mappings between input images and internal representations of machine vision systems (Cam-

marata et al. 2020). Any call to just "open the black box," however, is lacking, as the opacity of machine learning systems is an effect of complexity, not ideology. As Friedrich Kittler (2012, 163) puts it, the "entire hierarchical standard of self-similar program levels—from the highest programming language down to elementary machine code—rests completely flat on the material. In silicon itself there can be, to borrow from Lacan, no other of the other, which is also to say, no protection from the protection"—what we lack is not the ability to see what is going on, but to *interpret* it correctly.

The technical community, then, tends to open the black box but ignore the ideology, while perspectives from the humanities do the opposite, focusing only on cultural artifacts fed into (as training images) or extracted from (as generated images) the box. The missing component here is to see the black box itself as a cultural artifact. More generally, across all critical work in artificial intelligence, and especially within the emerging discipline of critical AI studies, we notice a tendency toward a kind of bifurcation between a position of extreme conservatism and one of absolute techno-futurism. A polarization that leads scholars to claim either that nothing has changed (contemporary AI systems are qualitatively no different to Babbage's calculating machines, Frank Rosenblatt's perceptrons, nineteenth-century IBM census machines,[25] and so forth); or that everything has (albeit the latter position is increasingly rare as, at least on the surface, it seems to mirror the corporate propaganda that is being critiqued in the first place).

Crucially, this polarization also exists at an epistemic level. Machine vision is either framed as "biased computational photography," fully determined by the latest politically flawed image dataset, or as an entirely unknowable, technically and epistemically opaque, alien technology. As we will argue, this polarization stems—at least in part—from a tension over the appropriate objects of study for a critical understanding of machine vision. Research that focuses on the visual culture of the training datasets of computer vision, of the sort we have illustrated above, necessarily inherits the short half-life of such datasets. On the other hand, intellectual histories

of computer vision can sometimes fail to give sufficient weight to contemporary technical shifts in computer vision algorithms, especially to those subtle differences in neural network architectures that have transformative epistemic consequences.

The proposal for a machine *Bildwissenschaft*—which we will begin to explore in the subsequent chapter—cannot circumvent these and other shortcomings. We are not raising the flag of a new discipline, but attempting to build a lens through which one of the opaquest technical phenomena of the past decades can be viewed slightly off-center. By reading these systems both technically and ideologically, we take seriously the intellectual legacy embedded in their design, while also questioning the epistemic ambitions they encode. In other words, we are trying to see artificial intelligence as an object worth studying, not merely as the latest manifestation of computation and its discontents, on its own terms—without defaulting to spectacle or disavowal.

Notes

Note: Unless otherwise noted, all translations are by the authors.

1 Consider, for instance, the following excerpt from an interview with the German philosopher Max Bense, from April 16, 1968. Bense is interviewed, for the television program *Aspekte*, about computer art. We hear the narrator state, "The kiss of the muse is cold. No feelings far and wide, only numbers count. But is the act of creation not the privilege of man? [. . .] If one aims to surgically remove, so to speak, the human factor from the creation of a work of art, one necessarily has to face the accusation that such a work would be literally inhuman. [. . .] All laws of beauty that could potentially be computed can only be based on the human-made works of art that already exist. Such laws would, then, only produce derivatives of what already exists, albeit in great variation." In other words, "Are AI images art?" rehashes precisely those concerns that had already been asked about the very first attempts at computer-facilitated image production. For a history of the Stuttgart School of computer art—to which Max Bense lent the philosophical framework—see, for instance, the academic work of Christoph Klütsch (e.g., 2007) and the curatorial work of Margit Rosen.

2 Obvious famously just executed code written by another artist, Robbie Barrat, to create *Edmond de Belamy.*

3 For a recent discussion of depth versus surface in the critique of artificial intelligence systems see Bajohr (2025) and Offert and Dhaliwal (2024).

4 We will use the term *machine vision* in this book, instead of the somewhat more common term computer vision, to emphasize, as will become apparent, the significance of technologies of machine learning as a medium that, while certainly determined by the affordances of computation, produces an independent set of media-theoretical and, by extension, sociopolitical questions.

5 The main training dataset for Stable Diffusion, LAION2B-en (a subset of LAION-5B), named for its *two billion images*, takes up 6.2 terabytes. It is thus over three thousand times larger than the abovementioned resulting model (though there are of course many versions and sizes of Stable Diffusion).

6 A byte, for reference, is what we use to store one pixel or one letter when we are being economical.

7 As recounted by Pliny the Elder: Zeuxis tricked birds into pecking on his lifelike grapes, but Parrhasius surpassed him by fooling Zeuxis himself, who, upon attempting to pull back the curtains on Parrhasius's picture, realized them to be painted.

8 The Turing Test is itself constantly misunderstood as an abstract benchmark, when in reality Turing gives us extremely clear guidelines for its implementation. "I believe that in about fifty years' time," Turing (1950, 442) writes, "it will be possible to programme computers, with a storage capacity of about 109, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning." In other words, around the year 2000, a computer with a storage capacity of roughly 120 MB should "win" the Imitation Game with a 70 percent chance of winning after five minutes of playtime.

9 More precisely—and perhaps more fairly to Turing—he states that the question of whether machines can think is intractable, but that the problem of whether machines can *appear* to think is the closest tractable problem.

10 Lovelace's algorithm (which included looping and branching) calculated the Bernoulli numbers.

11 One can find a variation of this point in the "The Argument from Consciousness" (451), where Turing argues that while computing machines *are* deterministic, they might still produce surprise, it might just look different from the "usual" surprise we expect from creative expression.

12 One of the most recent cases, picked up by many media outlets, was the "finding" that AI-generated poetry was preferred over human-written poetry (Porter and Machery 2024)—by people without much previous expertise or exposure to poetry, a limitation that both authors and reviewers failed to address.

13 Searle's research revolved around the problem of the speech act, first proposed by J. L. Austin, and the problem of phenomenological intentionality, a combination ideally suited to address the question of language-based artificial intelligence already raised in Turing's 1950 paper. His thought experiment became the focal point of the strong AI debate of the 1980s.

14 Imagine a machine, the argument goes, "so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we

should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for" (Leibniz 1898, 228).

15 The "Stochastic Parrots" paper mentioned above obviously stands in this tradition as well, though it cites neither Turing, Jefferson, Searle, nor Descartes.

16 As it is the case with Turing's machine-learning-like "child machines," discussed toward the end of his 1950 paper, the *viva voce* argument seems to anticipate not only the philosophical problems inherent in the attempt to build intelligent machines, but also some of the technical solutions emerging half a century later in response to them.

17 Searle (1980, 419) argues that he knows that the text "hamburgers" refers to hamburgers, whereas the computational "Chinese subsystem knows only that 'squiggle squiggle' is followed by 'squoggle squoggle.'"

18 Our proposed shift toward a closer analysis of the technical concreteness of artificial intelligence systems mirrors a shift that has long been completed in theoretical computer science, which now theorizes its own limits almost exclusively in terms of computational complexity (the difficulty of a problem in terms of the resources required to solve it) rather than computability (the difficulty of a problem in terms of its intrinsic tractability). In fact, research in computational complexity has shown that all fundamental philosophical problems within computer science—and many outside of it—are best described as problems of computational complexity (Aaronson 2011, 2013).

19 A traditionally underestimated function of media (Winkler 2015).

20 While Agre's work is still not quite prominent as a theoretical reference, the past five years or so have seen a resurgence of interest in it—not only around artificial intelligence, but also around surveillance. See, for instance, Bunz (2021).

21 More precisely, as we shall see in the following chapters: gradually since the 1980s, and quite rapidly during the contemporary turn toward multimodal foundation models.

22 Minsky and Papert (1988), famously, show how higher-complexity problems require multilayer perceptrons, but those—in 1969, so before backpropagation—could not be tuned efficiently.

23 Agre credits Ferdinand Braudel's book on everyday life in the Renaissance as the inspiration for the title of the paper "The Structures of Everyday Life" (1985).

24 Technically, ImageNet itself contains far more than one thousand categories. We are referring here to the most commonly used version of the dataset within computer vision applications, ImageNet-1k, sometimes also referred to as the ILSVRC-2012 subset. We should also remark here that one of the ironies of dataset critique is the outsized emphasis on outdated datasets that might have some sort of ghostly afterlife but do not inform models deployed currently.

25 Which they are in terms of their racism, but not functionally.

References

Aaronson, Scott. 2011. "Why Philosophers Should Care About Computational Complexity." Preprint, arXiv, August 8, 2011, 1108.1791.

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

Agre, Philip E. 1985. "The Structures of Everyday Life." Working Paper No. 267. MIT Artificial Intelligence Laboratory, February.

Agre, Philip E. 1997. *Computation and Human Experience.* Cambridge University Press.

Agre, Philip E. 2005. "The Soul Gained and Lost: Artificial Intelligence as a Philosophical Project." In *Mechanical Bodies, Computational Minds: Artificial Intelligence from Automata to Cyborgs*, edited by Franchi and Güven Güzeldere. MIT Press.

Agre, Philip E., and David Chapman. 1987. "*Pengi*: An Implementation of a Theory of Activity." In *Proceedings of the Sixth National Conference on Artificial Intelligence*, edited by John McDermott. AAAI Press / MIT Press.

Azar, Mitra, Geoff Cox, and Leonardo Impett. 2021. "Introduction: Ways of Machine Seeing." *AI and Society* 36, no. 4: 1093–104.

Barthes, Roland. 1981. *Camera Lucida: Reflections on Photography.* Translated by Richard Howard. Hill and Wang.

Baxandall, Michael. 1978. *Painting and Experience in Fifteenth-Century Italy: A Primer in the Social History of Pictorial Style.* Oxford University Press.

Bajohr, Hannes. 2025. "Surface Reading LLMs: Synthetic Text and Its Styles." Lecture given at the Center for the Humanities and Machine Learning at the University of California, Santa Barbara, January 24.

Bender, Emily M., and Alexander Koller. 2020. "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. John Wiley and Sons.

Benjamin, Walter. 1999. "Little History of Photography." In *Selected Writings 2, 1927–1934*, edited by Michael W. Jennings, Howard Eiland, and Gary Smith. Harvard University Press.

Berger, John. 1972. *Ways of Seeing*. Penguin.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." Preprint, arXiv, October 5, 2021, 2110.01963.

Bowker, Geoffrey, and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Brooks, Rodney. 1990. "Elephants Don't Play Chess." *Robotics and Autonomous Systems* 6: 3–15.

Bunz, Mercedes. 2021. "How Not to Be Governed Like That by Our Digital Technologies." In *The Ends of Critique. Methods, Institutions, Politics*, edited by Kathrin Thiele, Birgit M. Kaiser, and Timothy O'Leary. Rowman and Littlefield.

Buschek, Christo, and Jer Thorp. n.d. “Models All the Way Down.” Knowing Machines. https://knowingmachines.org/models-all-the-way.

Cammarata, Nick, Shan Carter, Gabriel Goh, et al. 2020. “Thread: *Circuits*.” *Distill*, March 10. http://www.doi.org/10.23915/distill.00024.

Chun, Wendy Hui Kyong. 2024. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition.* MIT Press.

Crawford, Kate, and Trevor Paglen. 2019. “Excavating AI: The Politics of Training Sets for Machine Learning.” https://excavating.ai.

Danto, Arthur C. 2013. *What Art Is.* Yale University Press.

De Vries, Terrance, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. “Does Object Recognition Work for Everyone?” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society.

Denson, Shane. 2023. “From Sublime Awe to Abject Cringe: On the Embodied Processing of AI Art.” *Journal of Visual Culture* 22, no. 2: 146–75.

Dobson, James E. 2025. “Beyond Computational Formalism, or Architecture Matters.” *Journal of Cultural Analytics* 10, no. 3 (September). https://doi.org/10.22148/001c.133923.

Descartes, René. 2006. *A Discourse on Method.* Edited by Ian Maclean. Oxford University Press.

Dick, Stephanie. 2015. “Of Models and Machines: Implementing Bounded Rationality.” *Isis* 106, no. 3: 623–34.

Dneprov, Anatoly. 1961. “Igra.” *Znanie–Sila* 5: 39–41.

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

Dreyfus, Hubert L. 2007. “Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian.” *Philosophical Psychology* 20, no. 2: 247–68.

Drucker, Johanna. 2011. “Humanities Approaches to Graphical Display.” *Digital Humanities Quarterly* 5, no. 1: 1–21.

Flusser, Vilém. 2013. *Towards a Philosophy of Photography. Translated by Anthony Mathews.* Reaktion Books.

Gaboury, Jacob. 2013. “A Queer History of Computing.” *Rhizome*, February 19. https://rhizome.org/editorial/2013/feb/19/queer-computing-1/.

Gaboury, Jacob. 2021. *Image Objects: An Archaeology of Computer Graphics.* MIT Press.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, et al. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64, no. 12: 86–92.

Gershgorn, Dave. 2022. “The Data That Transformed AI Research—and Possibly the World.” *Quartz*, July 20. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world.

Golumbia, Davis. 2024. *Cyberlibertarianism: The Right-Wing Politics of Digital Technology*. University of Minnesota Press.

Hooker, Sara. 2021. “Moving Beyond ‘Algorithmic Bias Is a Data Problem.’” *Patterns* 2, no. 4: 1–4.

Hoy, Meredith. 2017. *From Point to Pixel: A Genealogy of Digital Aesthetics.* Dartmouth College Press.

Impett, Leonardo. 2024. “Digital Art History as Critical AI.” *Art Bulletin* 106, no. 2: 11–14.

Impett, Leonardo, and Fabian Offert. 2024. "There Is a Digital Art History." *Visual Resources* 38, no. 2: 186–209.

Irani, Lilly. 2015. "The Cultural Work of Microwork." *New Media and Society* 17, no. 5: 720–39.

Jefferson, Geoffrey. 1949. "The Mind of Mechanical Man." The *British Medical Journal* 1, no. 4616: 1105–10.

Kataoka, Hirokatsu, Kazushige Okayasu, Asato Matsumoto, et al. 2022. "Pre-Training without Natural Images." *International Journal of Computer Vision* 130, no. 4: 990–1007.

Katz, Yarden. 2020. *Artificial Whiteness: Politics and Ideology in Artificial Intelligence.* Columbia University Press.

Kittler, Friedrich. 2012. "Protected Mode." In *Literature, Media, Information Systems.* Routledge.

Klütsch, Christoph. 2007. *Computergrafik: Ästhetische Experimente zwischen zwei Kulturen: Die Anfänge der Computerkunst in den 1960er Jahren.* Springer.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates.

Leibniz, Gottfried Wilhelm. 1898. *The Monadology and Other Philosophical Writings.* Translated by Robert Latta. Oxford University Press.

MacKenzie, Adrian, and Anna Munster. 2019. "Platform Seeing: Image Ensembles and Their Invisualities." *Theory, Culture and Society* 36, no. 5: 3–22.

Malevé, Nicolas. 2021. "On the Data Set's Ruins." *AI and Society* 36, no. 4: 1117–31.

Mansfield, Elizabeth. 2007. *Too Beautiful to Picture: Zeuxis, Myth, and Mimesis.* University of Minnesota Press.

Masís, Jethro. 2014. "Making AI Philosophical Again: On Philip E. Agre's Legacy." *Continent* 4, no. 1. https://philarchive.org/archive/MASMAP.

McQuire, Scott, Jasmin Pfefferkorn, Emilie K. Sunde, Celia Lury, and Daniel Palmer. 2024. "Seeing Photographically Seeing Photographically." *Media Theory* 8, no. 1: 1–18.

Menabrea, Luigi Federico. 1842. "Sketch of the Analytical Engine Invented by Charles Babbage. With Notes upon the Memoir by the Translator Ada Augusta, Countess of Lovelace." https://www.fourmilab.ch/babbage/sketch.html.

Meyer, Roland. 2020–24 X/Bluesky threads. https://bsky.app/profile/bildoperationen.bsky.social.

Meyer, Roland. 2021. *Gesichtserkennung.* Verlag Klaus Wagenbach.

Meyer, Roland. 2024. *Operative Porträts: Eine Bildgeschichte der Identifizierbarkeit von Lavater bis Facebook.* Konstanz University Press.

Minsky, Marvin, and Seymour Papert. 1988. *Perceptrons: An Introduction to Computational Geometry.* Expanded ed. MIT Press.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press.

Offert, Fabian. 2019. "The Past, Present, and Future of AI Art." *Gradient*, June 18. https://thegradient.pub/the-past-present-and-future-of-ai-art/.

Offert, Fabian. 2023. "KI-Kunst als Skulptur." In *KI-Realitäten. Modelle, Praktiken und Topologien des Maschinellen Lernens*, edited by Richard Groß and Rita Jordan. Transcript.

Offert, Fabian. 2024a. "Five Theses on the End of AI Art." In *Un/learn AI: Approaching AI in Aesthetic Practices.* https://unlearn.gestaltung.ai/article/pj83wpqd.

Offert, Fabian. 2024b. "KI-basierte Verfahren in der bildenden Kunst." In *De Gruyter Handbuch Künstliche Intelligenz und die Künste*, edited by Stephanie Catani. De Gruyter.

Offert, Fabian. 2025a. "Are Nearest Neighbors Good Neighbors?" In *The World Through AI: Exploring Latent Spaces*, edited by Antonio Somaini. JBE Books.

Offert, Fabian. 2025b. "The Raw and the Cooked: On Postdigital Painting in the Age of Artificial Intelligence." In *Between Pixel and Pigment: Painting in the Postdigital Present*, edited by Nina Gerlach and Simon Vagts. Reimer.

Offert, Fabian, and Ranjodh Singh Dhaliwal. 2024. "The Method of Critical AI Studies: A Propaedeutic." Preprint, arXiv, November 28, 2024, 2411.18833.

Offert, Fabian, and Thao Phan. 2024. "A Sign That Spells: Machinic Concepts and the Racial Politics of Generative AI." *Journal of Digital Social Research* 6, no. 4: 49–59.

Parikka, Jussi. 2023. *Operational Images: From the Visual to the Invisual.* University of Minnesota Press.

Pasquinelli, Matteo. 2019. "How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence." *Spheres* 5 (November). https://spheres-journal.org/contribution/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/.

Pereira, Gabriel, and Bruno Moreschi. 2023. "Living with Images from Large-Scale Data Sets: A Critical Pedagogy for Scaling Down." *Photographies* 16, no. 2: 235–61.

Porter, Brian, and Edouard Machery. 2024. "AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably." *Scientific Reports* 14, no. 26133. https://doi.org/10.1038/s41598-024-76900-1.

Raley, Rita, and Jennifer Rhee. 2023. "Critical AI: A Field in Formation." *American Literature* 95, no. 2: 185–204.

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. "Do ImageNet Classifiers Generalize to ImageNet?" In *Proceedings of the 36th International Conference on Machine Learning*, edited by Kamalika Chaudhuri and Ruslan Salakhutdinov. PMLR.

Richardson, Michael. 2024. *Nonhuman Witnessing: War, Data, and Ecology After the End of the World.* Duke University Press.

Roose, Kevin. 2022. "An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy." *New York Times*, September 2.

Salvaggio, Eryk. 2022. "How to Read an AI Image." *Cybernetic Forests*, October 2. https://cyberneticforests.substack.com/p/how-to-read-an-ai-image.

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–57.

Somaini, Antonio. 2023. "Algorithmic Images: Artificial Intelligence and Visual Culture." *Grey Room* 93: 74–115.

Somaini, Antonio, ed. 2025. *The World Through AI: Exploring Latent Spaces.* JBE Books.

Stability.ai. 2022. “Stable Diffusion Public Release.” August 22. https://stability.ai/news/stable-diffusion-public-release.

Steyerl, Hito. 2023. “Mean Images.” *New Left Review*, no. 140–41 (March–June): 82–97.

Travers, Michael. 2009. “Phil Agre, an Appreciation.” *Omniorthogonal*, November 21.

Turing, Alan Mathison. 1936. “On Computable Numbers, with an Application to the Entscheidungsproblem.” *Proceedings of the London Mathematical Society*, ser. 2, 42, no. 1: 230–65.

Turing, Alan Mathison. 1950. “Computing Machinery and Intelligence.” *Mind* 59, no. 236: 433–60.

Virilio, Paul. 1994. *The Vision Machine.* Translated by Julie Rose. Indiana University Press.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35: 24824–37.

Winkler, Hartmut. 2015. *Prozessieren: Die dritte und vernachlässigte Medienfunktion.* Fink.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books.

Zylinska, Johanna. 2020. *AI Art: Machine Visions and Warped Dreams.* Open Humanities Press.

Zylinska, Johanna. 2023. *The Perception Machine: Our Photographic Future Between the Eye and AI.* MIT Press.

[2]

Machine *Bildwissenschaft*

> For human vision to be explained by a computational theory, the first question is plain: What are the problems that the brain solves when we see? It is argued that vision is the construction of efficient symbolic descriptions from images of the world.
>
> —David Marr, "Visual Information Processing: The Structure and Creation of Visual Representations"

Disembodied Vision

What is the relationship between seeing and knowing?[1] In early 2024, OpenAI released SORA, a computer vision model apparently capable of producing amazingly detailed videos from text prompts. The examples were cherry-picked, of course, and shown for only about ten seconds, but they were also clearly qualitatively superior to anything that might have been possible with previous video-generation systems. This generated huge public interest—understandably, as the same lab had revolutionized image generation three years earlier with their DALL·E and CLIP "multimodal" text-image networks.

The costs associated with commercial video production are orders of magnitude greater than those for the production of still images, photorealistic or otherwise. Thus the relatively minor economic consequences (at least so far) of neural image generation may well be dwarfed by the effects of cheaply synthesizing photorealistic video. Or, indeed, by the mass production of 3D models for computer graphics and video games, still a costly, time-consuming, skilled task. Neural video synthesis models can be easily adapted to generate such "assets," since they tend to have good models of three-dimensional form, depth, and space.

[Figure 2.1]. Metapictures; still from a SORA-produced video with the prompt "The camera rotates around a large stack of vintage televisions all showing different programs—1950s sci-fi movies, horror movies, news, static, 1970s sitcom, etc., set inside a large New York museum gallery." See "Creating Video from Text," OpenAI, February 15, 2024, https://openai.com/index/sora/.

The technical paper[2] that presents SORA, however, is relatively uninterested in Hollywood. Its title gives away the radical claim: *video generation models as world simulators.* Realistic video generators are far more than just a technical extension of image generators,[3] according to this argument—they must also understand *dynamics*, including but not limited to three-dimensional structure, basic physical interactions (solid objects cannot pass effortlessly through each other), object permanence (objects rarely seem to appear or disappear into nothing), and even some of the laws of physics, especially gravity and the optics of light and shadow (though probably not nonvisible radiation or angular momentum). Video data implies *causality*, the argument goes, and a system capable of guessing what happens next in a given shot, must have at least some basic understanding of the physical world. Building better video models, according to OpenAI, is "a promising path towards the development of highly-capable simulators of the physical and digital world, and the objects, animals and people that live within them" (Brooks et al. 2024). Far from replacing Hollywood studios, OpenAI seems to see the real power of video synthesis

as the formation of a kind of hallucinatory training ground for autonomous robots, in which other machine learning models can learn much faster than real time.[4]

This was, of course, a controversial claim,[5] rejected, for instance, by Yann LeCun, formerly chief AI scientist at Meta (formerly known as Facebook) and himself a key figure in the history of computer vision and machine learning.[6] Not that LeCun doesn't believe in world models—he has himself outlined a meta-architecture for a quite complex world model (LeCun 2022). Rather, he doesn't believe in the continuity between "realistic" visual generation and physical, tactile knowledge about the outside world. In an interview in late 2023, just before the release of SORA, LeCun (2023) argued that world models should be primarily abstract models of causality, and thus "don't need to predict all the details about the world, they don't need to be generative, they don't need to predict exactly every pixel in a video." Four days after SORA's release in February 2024, he reiterated his opposition on a social media platform: "Modeling the world for action by generating pixel [*sic*] is as wasteful and doomed to failure as the largely-abandoned idea of 'analysis by synthesis.' [. . .] That's why generative models for sensory inputs are doomed to failure."

The Modern Molyneux

By way of example, LeCun asks us to imagine a robot being asked to screw together planks of wood. With a perfect video model, the robot would be able to reproduce a realistic wood grain and imagine *this* specific screwdriver down to the last pixel. But, he asks, could it be taught to screw together planks of wood by video alone, without ever moving its own arms?

OpenAI and Meta, here, seem to be rehearsing very precisely a debate posed in 1688 by the Irish philosopher William Molyneux (1996, 17) to his friend John Locke:

> A Man, being born blind, and having a Globe and a Cube, nigh of the same bignes, Committed into his Hands, and

> being taught or Told, which is Called the Globe, and which the Cube, so as easily to distinguish them by his Touch or Feeling; Then both being taken from Him, and Laid on a Table, Let us Suppose his Sight Restored to Him; Whether he Could, by his Sight, and before he touch them, know which is the Globe and which the Cube? Or Whether he Could know by his Sight, before he stretch'd out his Hand, whether he Could not Reach them, tho they were Removed 20 or 1000 feet from Him?

Molyneux's problem attracted such interest—not just from Locke, but also from George Berkeley, Voltaire, Denis Diderot, Gottfried Wilhelm Leibniz, and Hermann Ludwig Ferdinand von Helmholtz—that, for the philosopher Ernst Cassirer (1955, 108), it constitutes the central question around which the history of Enlightenment psychology and epistemology can be told.

It was no coincidence that the thought experiment was first conceived in a letter to Locke, or more precisely to the "Author of the *Essai Philosophique concernant L'Entendement*." There, Locke sets out a theory of knowledge entirely compatible with the claims of machine learning: that no human knowledge is innate, but that it derives entirely from experience, which is to say, from either perception ("sensation") or reflection. Thus a blind man, in Locke's example, could never have a notion of color.

Locke's example stands in a long tradition here that extends in both directions. The philosophy of artificial intelligence—inheriting, of course, from a more general philosophy of perception—has been structured, throughout its existence, by a peculiar relationship to the philosophical and ethical questions raised by disability. This begins already with Descartes's famous 1637 discussion of intelligence in chapter 5 of his *Discourse on the Method*, where two kinds of disability serve as proof for the exceptionality of the mind: on the one hand, the men "born deaf and dumb" who still can communicate like humans, albeit by using not speech but signs; and on the other, the "child with a disordered brain" (Descartes 1998, 32) who still surpasses any animal in its capability to communicate.

The instrumentalization of disability persists in contemporary philosophical discourse on artificial intelligence. Howard Robinson's (1982, 4) classic thought experiment, formulated to support the knowledge argument against physicalism, notably employs a scientist with a disability to illustrate the nature of qualia. Disability is thus repeatedly deployed as a conceptual device that tests the boundaries of perception and cognition. As Thomas Nagel (1974, 442) observes, the "distance between oneself and other persons and other species can fall anywhere on a continuum," and within the human species, the closest point of comparison is often a person with a perceptual disability who compensates through alternative means. Even for Descartes, writing in the early seventeenth century, sensory impairment did not challenge the essence of humanity, since he maintained a binary distinction between possessing a soul and lacking one. In contrast, for Nagel, while a continuum exists, its extremes merely delineate the objective gap between subjective experiences.

Disability, much like gender, has historically been appropriated as a source for "edge cases"—an epistemic tool through which the limits of intelligence, perception, and personhood are conceptually tested.[7] Molyneux's problem helps us trace the intellectual history of questions in visual epistemology that have been contested throughout the development of computer vision by asking: Is our visual perception fundamentally anchored to a more basic sense, e.g. of touch? At the same time, it serves to illustrate the ways in which the fundamental "technical roadblocks," to speak again with Agre, of artificial intelligence research are always tied to such "edge cases" because, essentially, they are all epistemic problems: they are problems about the limitations of knowledge and its making.

An alternative is emerging in disability hacktivism and contemporary disability studies, where disability is not treated as a limit to be tested, but as a site of knowledge (Alexander 2025): a situated, embodied standpoint from which to rethink norms of perception, interaction, and intelligence, surfaced through "moments of technofailure" (Alexander and Sterne 2024, 188). We see this, for

instance, in impairment phenomenology (Sterne 2021) and in the broader practices of crip technoscience (Hamraie and Fritsch 2019), which refuse the abstraction of disability into metaphor or thought experiment and instead center it as a generative condition for critique.

Are Pictures Vision?

Berkeley, writing on Molyneux's problem at the start of the eighteenth century, argued that there is no innate relation between vision and touch, between the world of images (which he called *visible ideas*) and the world of three-dimensional physical structure (*tangible ideas*). Their relation is entirely arbitrary, and during development we must learn to correlate visual signals with touch in order to be able to judge distances, sizes, and so forth. In what we might anachronistically call a radically multimodal (in the computer science sense) theory of perception, Berkeley uses language as a metaphor for vision. This account of vision was extraordinarily long-lasting. By the mid-nineteenth century, John Stuart Mill described it as "one of the least disputed doctrines in the most disputed and most disputable of all sciences" (Atherton 1990, 4).

For Berkeley, we decode visual forms into three-dimensional tactile structures in much the same way as we learn to interpret the sound of speech into some kind of semantic meaning. As Voltaire (1738, 68) later paraphrased, "We learn to see, exactly as we learn to speak and read." There is an implicit hierarchy here, preempting theorists of computer vision such as David Marr, in which three-dimensional geometric structure is the hidden semantic content of vision. It follows that the purpose of vision is to reconstruct the distances and locations of objects in the world.

What might Berkeley have thought of a model like SORA—of OpenAI's wager that a sophisticated visual intelligence need not have an actual body in order to *model* the world? Take the "Case of an Intelligence, or Unbody'd Spirit, suppos'd to see perfectly well [. . .] but to have no Sense of Touch," writes Berkeley in 1709,

at the end of the *New Theory of Vision*—though "Whether there be any such Being in Nature or no, is beside my Purpose to enquire" (2012, 64). Without a sense of touch, such an intelligence could never have an idea of three-dimensional solidity, nor distance, "nor consequently of Space or Body." Even two-dimensional geometry, argues Berkeley, would be beyond our eighteenth-century artificial vision system, since the "Intelligence" would have no use for a ruler or a compass.[8]

Less than fifty years after Berkeley, the French philosopher Étienne Bonnot de Condillac gave an even more radically robotic analogy to address Molyneux's problem: imagining, quite literally, a kind of automaton (referred to as a statue) in place of the blind man, which gradually acquires each of the five senses. Condillac proposes an even more extreme version of Locke and Berkeley's perspective: Not only can a machine with only vision not discern shapes or volumes, it is even blind to distances and locations—it sees the world as we see a difficult painting "whose subject is unfamiliar." Condillac (2014, 214) thus attributes a special difficulty to vision—touch, for instance, is far more self-sufficient. He supposes that the machine would eventually have access to all kinds of mental processes simply from perception: first attention, then pleasure, pain, memory, comparison, surprise, ideas, imagination. All these arise, fundamentally, from perception and *through* pleasure and pain. Condillac's argument, in that, is very close to that of an influential 2021 paper from Google's DeepMind, which argues that reinforcement learning—wherein an algorithm learns from being "rewarded" as a result of behavior that should be encouraged—is a sufficient basis for a much wider set of behaviors, which computer scientists call *artificial general intelligence* (Silver et al. 2021).

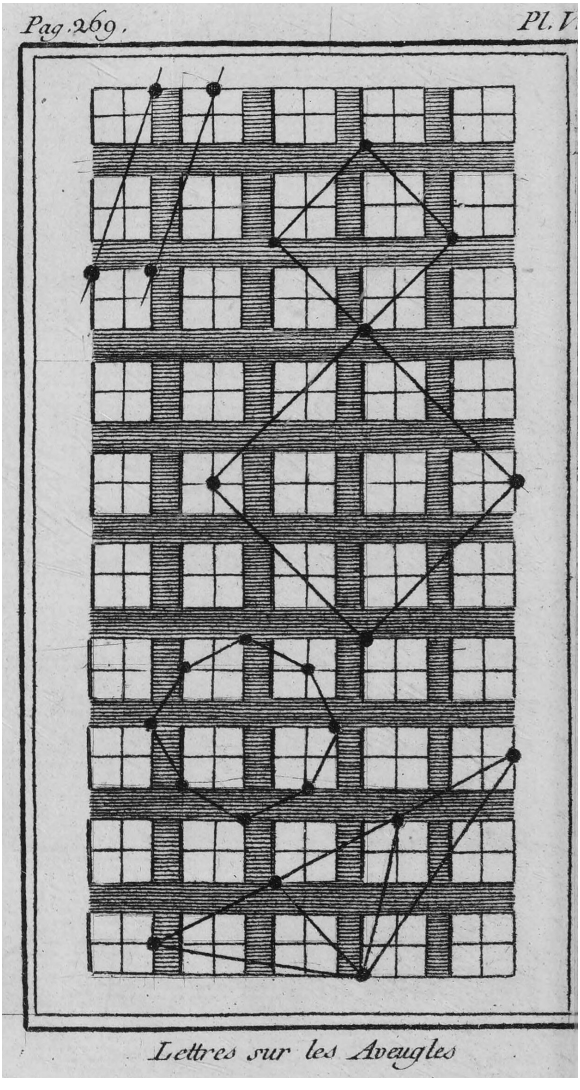
Locke, Berkeley, and Molyneux's views on the problem appeared to be confirmed when, in 1728, the English surgeon William Cheselden published an account of cataract removal surgery on a thirteen-year-old boy who had been blind from birth. The boy, who we now know to be Daniel Dolins (Leffler et al. 2021), struggled immediately with even the most basic aspects of visual perception.

Cheselden (1728, 449) reports Dolins's particular difficulty with two-dimensional pictorial representations:

> We thought he soon knew what Pictures represented, which were shew'd to him, but we found afterwards we were mistaken; for about two Months after he was couch'd, he discovered at once, they represented solid Bodies; when to that Time he consider'd them only as Party-colour'd Planes, or Surfaces diversified with Variety of Paint; but even then he was no less surpriz'd, expecting the Pictures would feel like the Things they represented, and was amaz'd when he found those Parts, which by their Light and Shadow appear'd now round and uneven, felt only flat like the rest; and ask'd which was the lying Sense, Feeling, or Seeing?

Cheselden's observations on two-dimensional representations—an aspect absent from Molyneux's original question—sparked a broader debate on what we will call the *picture question* of vision: the relationship between visual perception and pictorial representation. While this debate was informed by earlier seventeenth-century discussions on the retinal image, demonstrated by Johannes Kepler and extensively theorized by Descartes, it remained distinct in its implications. Ultimately, the picture question would emerge as the defining fault line in the philosophy of vision underpinning the development of computer vision.

Denis Diderot took up the case of Cheselden's patient and the picture question in his *Lettre sur les aveugles à l'usage de ceux qui voient* (*Letter on the Blind for the Use of Those Who See*), published in 1749. Through a discussion of the experiences of the blind Cambridge mathematician Nicholas Saunderson[9]—including a detailed examination of his tactile instruments for solving mathematical problems, a form of highly embodied computation—Diderot explores visual perception and the limits of sensory experience. This inquiry led him to a kind of atheist metaphysics, a stance that ultimately caused his imprisonment in the Château de Vincennes a month



[Figure 2.2]. A diagram of the “palpable arithmetic” of Nicholas Saunderson, from Diderot (1772, plate 5).

after the book's publication. He compares Dolins's confusion over pictures to similar experiences from non-European peoples: "[The boy] was vastly surprised at finding a plane surface without any relief. . . . Painting likewise has the same effect on savages. They take the painted figures for living men, question them and are astonished at receiving no answer; and this error in them certainly did not proceed from their not being accustomed to see" (Diderot 1749, 161).

This shift moves the picture question beyond sensory impairment to broader epistemological and cultural concerns. No longer confined to the perceptual challenges of the blind, Diderot extends the argument to the ways colonized people encounter and interpret pictorial representation—reinscribing a hierarchy in which both disability and colonial (and thus racial) difference test the boundaries of European theories of vision.

Cheselden and Diderot, then, are proposing that pictures are a kind of limit case of thinking about vision. If, as Locke, Berkeley, and Molyneux had suggested, visual perception is primarily a learned skill—that is, if visual information is already at least partially arbitrary—then pictures are a convention within a convention. "Every image embodies a way of seeing" (Berger 1972, 10), and thus an implicit theory of vision, making visible the assumptions that govern perception, representation, and interpretation. Pictures, then, are always what W. J. T. Mitchell (1994) calls metapictures: images that not only represent but also reflect on the nature of representation itself. Whether by containing other images, recontextualizing visual forms, or serving as exemplars of pictoriality, metapictures, in Mitchell's formulation, expose the processes through which vision and meaning are constructed. As we shall see, the precise mechanics of this relationship between pictures and vision—between perception and representation, illusion and convention—would come to define the central debates in the philosophy of computer vision, and indeed artificial intelligence more widely.

Brooks and Embodied Robotic Vision

The debate over the learning potential of abstract intelligences with vision *alone* (i.e., without a tangible or tactile body) was reignited in the last decades of the twentieth century with a new wave of theoretical and technical reflections on the importance of embodiment in artificial intelligence, and especially in computer vision. In the late 1970s and 1980s, Hans Moravec was already experimenting with robotic rovers equipped with TV cameras that could slowly navigate a room—first at Stanford, a relic of moon rover research in the 1960s, and then at Carnegie Mellon's Robotics Institute, which itself became a center for autonomous vehicles, controversially largely taken over by Uber in 2015. In what is now known as the Moravec paradox, he observed that while it is comparatively easy to teach a computer accounting or chess, it is "difficult or impossible to give them the skills of a one-year-old when it comes to perception" (Moravec 1988, 15). Once again, artificial intelligence is examined through the lens of alterity—not through disability, gender, or race, but through the developmental limitations and capabilities associated with early childhood.

Drawing on Moravec's work, the MIT roboticist Rodney Brooks argued, in his seminal paper "Elephants Don't Play Chess" (1990), that AI research had been wrong to reduce intelligence to the manipulation of symbolic information. For Brooks this position, which he calls the symbol system hypothesis, spawned the mistaken attempt to produce computer vision systems that would describe images in terms of symbolic information. Instead, Brooks and his group sought to create machine intelligence that was grounded in its physical context, famously declaring that the world is its own best model (the physical grounding hypothesis). He argued that attempts to build and continuously update an internal computational model of the outside world were intractable, and branded his alternative approach Nouvelle AI. Many of Brooks's early prototypes approximated the behavior of insects or simple vertebrates. Phil Agre completed his PhD in Brooks's lab at the end of the 1980s, working on Nouvelle AI approaches to planning.

In a guest editorial in the November 1991 *International Journal of Computer Vision*, Randal C. Nelson outlined the principles of a radically embodied, *behavioral* approach to computer vision based on Brooks's vision for embodied AI. A weak version, argued Nelson, is that some classical problems of computer vision that have proven difficult can be effectively learned by interacting with the world; a stronger version is that these problems are themselves ill posed, and that better ones might be found by looking at animal vision, starting not with humans but with the evolutionarily simplest forms of vision. Nelson points several times to a theory of vision that could form the foundations for such a move in computer vision: the "animate, behavioral, or active vision" (Nelson 1991, 5) of American psychologist James J. Gibson.

Gibson's Ecological Visual Perception

Gibson is one of the two visual psychologists of the twentieth century on whom computer vision has drawn most explicitly. He devoted the first page of his *The Perception of the Visual World* (1950) to a quote from Berkeley—with whom he credits various important discoveries on the relation between touch and sight, the importance of learned sensation (with Locke), and the construction of "visual space." Thanks to his prior work on the psychology of automobiles, Gibson had worked during the Second World War with the U.S. Army Air Force on visual aircraft identification[10] and pilot training films (Gibson 1982), specifically in the Motion Picture Unit of the Aviation Psychology Program, with access to Hollywood studio technicians. From this work came two preoccupations that defined Gibson's account of visual perception.

First, the importance of *motion* in vision. For instance, a pilot landing a plane does not recognize a series of still images of a runway, which gradually increase in scale, but rather perceives directly the runway getting larger (i.e., closer) in their field of vision. Furthermore, the expansion of one point in the horizon is perceived not as the motion of the world, but as egocentric movement: We are moving through the world and not the other way

around. Sometimes we can be tricked by this—as, when waiting for a train to depart, we see the train on the platform opposite start to move, and for a second wrongly assume we are moving. Gibson's idea of an "optic flow" (now usually called *optical flow*) turned out to be a practical principle for computer vision, which in the last decades of the twentieth century was still more heavily linked to robotics labs (and thus embodied applications) than it is today. A great many mathematical techniques were proposed for estimating optical flow (many of which were rules-based, and had no need for machine learning), either sparsely (tracking points through time) or densely (measuring the visual "speed" of every pixel in a video). These could lead to motion estimation, but also to 3D reconstruction ("structure from motion") and even video compression. Gibson thus stressed the importance of vision's ecological and biological context, of vision as a sense evolved to search for food, circumnavigate obstacles, avoid predators, and so forth. He also stressed the importance of the visual information not as "object detection," but rather as *affordances* (a term that has become influential in design thinking): Thus a fire says, "Don't touch"; tall cliffs say, "Careful not to fall over"; chairs say, "Sit on me," and so on.

Though Gibson's work was to become influential in computer vision, especially within robotics, he explicitly rejected the idea that visual perception could be modeled in terms of Claude Shannon's information theory. Shannon's model was based in radio communication—purposeful communication through a line with a finite transmission capacity—and this, for Gibson (1979, 242), could never be a model of visual perception: "The world does not speak to the observer. [. . .] Words and pictures convey information, carry it, or transmit it, but the information in the sea of energy around each of us, luminous or mechanical or chemical energy, is not conveyed. It is simply there. [. . .] [When] neuropsychologists began thinking of nerve impulses in terms of bits and the brain in terms of a computer, the applications did not work."[11] Gibson's wager was thus that visual perception works "directly" (closer to the eighteenth-century Scottish philosopher Thomas Reid than to

David Hume or Berkeley), that there is no internal mental image of the outside world formed. Rather, important visual phenomena (e.g., things getting larger) exist without the mediation of an internal retinal mental image. Gibson thinks we have been too distracted by Kepler's discovery of the "retinal image." It is unnecessary to explain visual perception, he argues, as the compound eyes of insects work perfectly well without ever containing an optically projected physical image. "The deep-seated notion of the retinal image as a still picture has been abandoned" (Gibson 1979, 238).

Building on insights from his wartime research on visual perception, Gibson introduced a pivotal distinction—one that went beyond merely challenging the concept of the retinal image. He underscored the fundamental difference between perceiving a physical scene and perceiving a picture, a distinction of particular significance given that both psychological research and the military training materials he contributed to had traditionally relied on photographs as proxies for direct visual experience. However, for Gibson, such images do not even constitute a first-order approximation of the visual field they purport to represent. In his most influential work, *The Ecological Approach to Visual Perception*, he dedicates the final section to the topic of depiction, including one chapter outlining a theory of drawings and another exploring a theory of cinematic representation.

Gibson felt the need to establish an entirely separate account of the psychology of the perception of pictures, precisely because he held that it was a mode of perception alien to our conventional "ecological" vision: "A picture is not like perceiving" (Gibson 1978, 231). Gibson is thus suspicious of the retinal image, the mental image, *and* the physical image.[12] This lifelong interest in pictures intensified over the last few decades of his life, surely at least in part thanks to his friendship with the art historian Ernst Gombrich, who spent seven years as Gibson's colleague at Cornell. Gibson and Gombrich cite each other frequently in the decade leading up to Gibson's death in 1979, and had a sort of public debate through an exchange of letters in *Leonardo* magazine. Gombrich was

suspicious of Gibson's too-clean separation between reality and pictures and cites false teeth and plastic flowers as examples that illusions can also be three-dimensional, and the relief as the form that shows the continuity between the two: from a few millimeters of depth (say, the head of a coin) to something like sculpture. They also disagreed on *how* a picture is made: Gombrich asked Gibson whether a painter making an image of Zermatt would first have to circumnavigate the mountain to perceive its "invariants," a point Gibson ultimately conceded.

The Picture Theory of Vision

While Gibson and Gombrich debated the nature of depiction and its relationship to perception, a parallel but distinct approach to vision was emerging from the field of computational neuroscience. David Marr had trained in mathematics and then theoretical neuroscience at Cambridge, working on information organization, memory, and learning in the brain.[13] He moved to the Artificial Intelligence Lab at MIT in 1973, where he would work on computational approaches to vision until his death at age thirty-five in 1980—just a year after James Gibson's.

Marr's greatest influence comes from a book he had worked on in the last year of his life, published posthumously in 1982: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* The opening to the first chapter makes Marr's view on human vision quite clear. The "process of discovering what is present in the world, and where it is" is "first and foremost, an information-processing task" (Marr 1982, 3). Understanding vision, he argues, requires understanding both how to extract the relevant information from images, and how to represent it internally in the brain so that it might be useful for whatever the organism has evolved to attend to (thus a spider, a pigeon, and a human might have quite different representations of the same image). Marr's approach is thus shaped by what he calls the duality of information processing and information representation.

Whereas Gibson had rejected a Shannonian account of vision as an information-theoretical task, Marr takes this as his central axiom: The dynamics of vision are dictated by information extraction and the creation of useful internal representations. Marr thought (as did Gombrich) that Gibson's "direct" account was overly optimistic about the difficulty of actually extracting such things as invariants from image fields. Many phenomena that Gibson had labeled direct perception had turned out to be rather difficult to capture computationally—Marr (1982, 31) describes Gibson as "misled by the apparent simplicity of the act of seeing." Nevertheless, for Marr, Gibson was asking the right questions. The first chapter of Marr's book, on the philosophy of his approach—the part of Marr's work which continues to be most influential—discusses Gibson at length, and makes it quite clear that his account of vision is "perhaps the nearest anyone came to the level of computational theory" (Marr 1982, 29), the kind of theory Marr himself was attempting to build.

Just as important as Marr's theory of vision and perhaps more long-standing is the way that theory was framed. The point of a computational theory is not merely to build automatic algorithms for vision, but to have access to a language in which the functions of parts of the visual system can be described. For Berkeley and Voltaire, seeing is like speaking or reading. Marr's sole example of a properly computational theory is Noam Chomsky's account of language—though he admits it could not actually be implemented on a computer. Instead, such a theory sits at the base of three hierarchical levels of explanation, which Marr developed with Tommaso Poggio:

1. Computational: What is the goal and rationale of the information processing task?
2. Algorithm and representation: How could it be implemented? What is the representation for the output and the algorithm to go from input to output?
3. Hardware: How can the algorithm and representation be realized physically?

Marr and Poggio's three-level framing of vision (and cognition more widely) far outlasted any of Marr's specific claims about visual neuroscience; it has been described as "one of the most well-known theoretical constructs of twentieth-century cognitive science" (Willems 2011). Take Gibson's notion of optical flow, for instance: Gibson gives us the goal of the task (to recognize moving textures), which can be calculated and represented computationally in various ways (for instance, sparsely and densely), and each of those ways might be implemented on different bits of hardware. But if Gibson saw his work as rejecting "the picture theory of natural perception," David Marr (1982, 31) was to be its greatest modern proponent: "Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information. [. . .] We have already seen that a process may be thought of as a mapping from one representation to another, and in the case of human vision, the initial representation is in no doubt—it consists of arrays of image intensity values as detected by the photoreceptors in the retina." There is a curiously ekphrastic logic here: Vision is the *description* of images, and therefore the description of pixels ("image intensity values"). In the Marrian framework, the ultimate function of vision is to produce a kind of internal 3D map (after an intermediate "2.5D sketch"), a paradigm very much maintained within research on autonomous driving. In this account, computer vision is effectively the inverse problem to computer graphics, where images are made from detailed 3D models. This paradigm, still influential in computer vision research, is known as *vision as inverse graphics.*[14]

But what happens when we look at a photograph we are holding in our hand? Marr's theory of vision here experiences a kind of short circuit. We see depth in the picture—Marr here takes after the Renaissance perspectival theorist Leon Battista Alberti, for whom every picture is a window onto an imagined space.[15] But the depth in the picture is clearly contradicted by the physical flatness of the image surface, which lacks the dimensionality of the space it depicts. Nonphotographic images with a more complex relation-

ship to perspective—self-contradictions and ambiguities—move even further outside the boundaries of what a Marrian account of vision can easily accommodate.

Marr's framing is often described as "vision as information processing." We might suggest that "vision as picture processing" is a more telling characterization. But this framing invites a more fundamental question: What kind of theory of pictures implicitly underpins computer vision models? The datasets used for training—be they collections of driverless car images, scans of receipts, or satellite photos—reflect unique properties and constraints of photographic media. These constraints are not mere technical parameters. If we take seriously, following Gabriel Pereira and Bruno Moreschi, the image dataset as a kind of photographic archive, then trained models constitute a latent theory of photography. A model tuned on a dataset from urban street scenes, for instance, will develop a vision tuned to particular visual affordances, such as the navigation of roads or recognition of street signs.

If neural models of vision inherit the stylistic and ideological dimensions of the photographic archives (i.e., datasets) on which they are trained, then a model of historical culture is, unavoidably, also a model of the photographic conventions that define it. Models like CLIP and DALL·E, when tasked with recalling images of fascism, consistently reproduce its characteristic black-and-white, high-contrast tones. Not merely out of technical convenience, but because these tones have come to signify the 1930s within the model's learned visual vocabulary. In this way, the model's understanding of history is bound to a photographic grammar that naturalizes a single mode of representing this era, sanitizing it by fixing it at a distance in a "media prison" in which historical events are not just represented, but visually constrained by a latent theory of the medium (Offert 2023). Thus, vision models both inherit and impose a particular media theory that dictates how the past may be viewed, understood, and ultimately constrained by the conventions of the medium itself.

Theories of visual perception, then, pivot on two fractures: first, the relationship between vision and embodiment, how sight aligns with or diverges from touch and physical presence; and second, the immediacy of vision, questioning whether seeing is an innate faculty or learned through experience. These fractures collide in the realm of pictures, which encompass both extremes: They are deeply learned forms of representation, yet offer a mode of vision that seems removed from tangible interaction. Though, in truth, the relationship to touch is far more complex: from digital pictures that lack physicality to heavily textured paintings or bas-reliefs that invite, even simulate, a tactile experience, as in Bernard Berenson's (1896) "tactile values" or David Freedberg's (2017) account of bodily empathy.

Computer vision, shaped in Marr's shadow, rests on an unstated assumption of continuity between vision and pictures, overlooking that it is fundamentally modeling a medium (photography), perhaps even a genre (amateur smartphone photography), rather than a sense (vision). Computer vision's latent theory of photography, as Malevé and Sluis (2023) put it, stems from precisely these "ontological conflations between 'real world' and photograph, human eye and camera, snapshot and visual stimulus." Or, we might add, OpenAI's conflation of TikTok-slop video generation and world simulation. Conceptual or even philosophical contradictions that, as Agre suggested, tend to surface precisely in AI's highly material, *technical* impasses.

Notes

1 This chapter owes a large debt to the connections between the history of philosophies of vision (especially those of the European Enlightenment) and computer vision first laid out by the art historian Michael Baxandall (1995, 60), who described neural networks as a "plastic, reactive, association machine that would have pleased Condillac more than Locke."

2 OpenAI's use of that term initially described something close to a *technical report*, a kind of self-published scientific paper, typeset in LaTeX and hosted on arXiv; it now describes something between a blog post and a product launch page.

3 It is difficult to overstate the genuine surprise felt by the technical community

on seeing the first results of GAN-based image generation in the late 2010s, and diffusion models shortly thereafter; but some sort of realistic image synthesis was, under restricted conditions, possible even with early-1990s linear techniques such as eigenfaces.

4 This framing of video models as "world simulators" marks a partial return to the original ambitions of computer vision, which was long entangled with robotics and questions of embodied perception (cf. Barlow 1961; Marr 1982). In those contexts, vision was understood as situated and sensorimotor—part of a physical feedback loop between organism and environment. The rise of deep learning, by contrast, has largely decoupled vision from embodiment, reconfiguring it as a pattern-recognition task performed over vast, disembodied image datasets. We will return to the "picture theory of vision" later in this chapter.

5 For an excellent investigation into the technical validity of the claim, see Millière (2024).

6 Both as the developer of the first convolutional neural network with backpropagation, LeNet, and later as a pioneer of genuinely deep neural networks for text in the 2010s. LeCun announced in November 2025 that he would leave Meta at the end of the year to found a startup around world models; the departure also appears to be due to his skepticism towards ever-larger language models as a pathway to so-called artificial general intelligence.

7 For the significance of blindness particularly in art history, see, for instance, Bexte (1999).

8 Berkeley does admit one other possibility. The "Intelligence" still sees different colors and intensities of light at every point in the visual field (or pixels, in case you've forgotten this book is about computers); Berkeley (2012, 65) speculates that one might accurately compute their magnitude, but light intensities are so variable and inconsistent that even "if we suppose it possible to be done, [it] must yet be a very trifling and insignificant Labour." His instinct, of course, was right—computer vision is hard.

9 Saunderson was possibly the first to formulate what became known as Bayes's theorem, the fundamental formula of probability that underpins most contemporary probabilistic machine learning.

10 See Irwin (2024) for a history of the entanglement of pattern recognition and military technology during the mid-twentieth century.

11 Later in the same chapter, he accuses these computationally inspired accounts of vision of being on the bandwagon (a term used in a famous paper by Shannon himself): "But it seems to me that all they are doing is climbing on the latest bandwagon, the computer bandwagon, without reappraising the traditional assumption that perceiving is the processing of inputs" (Gibson 1979, 251).

12 Gibson (1979, 285) cites Panofsky, calling perspective "symbolic," and thus, a bit like a language, arbitrary. This is perhaps a misreading of Panofsky, who is using Cassirer's formulation of symbolic form. Despite acknowledging the ambiguity of perspective, Gibson ultimately rejects the idea that pictorial systems can be freely invented in the same way as linguistic structures.

13 Specifically in the cerebellum, hippocampus, and neocortex, the last of which is motivated by David Hubel and Torsten Wiesel's work on feature detection

in the visual cortex. Marr's earlier work attempted to show that "the brain's central function is statistical pattern recognition and association, carried out in a very high-dimensional space of 'elemental' features" (Edelman and Vaina 2001, 596).

14 Marr himself seems not to use the term *inverse graphics,* which comes at least from Baumgart (1974) and has since become a paradigmatic term in computer science.

15 Computer vision often assumes this, but—as Panofsky showed—this is actually a cultural convention, a symbolic form, with no relation to what's going on in the retina whatsoever.

Bibliography

Alexander, Neta. 2025. "From the Mean Image to the Kin Image." Keynote address given at the conference Ethics and Aesthetics of Artificial Images, Venice, May 8–10.

Alexander, Neta, and Jonathan Sterne. 2024. "Beyond Access: Transforming Ableist Techno-Worlds." In *Technics: Media in the Digital Age*, edited by Nicholas Baer and Annie Oever. Amsterdam University Press.

Atherton, Margaret. 1990. *Berkeley's Revolution in Vision.* Cornell University Press.

Barlow, Horace. 1961. "Possible Principles Underlying the Transformation of Sensory Messages." *Sensory Communication* 1, no. 1: 217–33.

Baumgart, Bruce Guenther. 1974. *Geometric Modeling for Computer Vision.* Office of Naval Research ARPA, Stanford University.

Baxandall, Michael. 1995. *Shadows and Enlightenment.* Yale University Press.

Berenson, Bernard. 1896. *The Florentine Painters of the Renaissance.* G. P. Putnam's Sons.

Berger, John. 1972. *Ways of Seeing.* BBC and Penguin.

Berkeley, George. 2012. *An Essay Towards a New Theory of Vision.* In *Berkeley: Philosophical Writings*, edited by Desmond M. Clarke. Cambridge University Press.

Bexte, Peter. 1999. *Blinde Seher: Wahrnehmung von Wahrnehmung in der Kunst des 17. Jahrhunderts.* Verlag der Kunst.

Brooks, Rodney. 1990. "Elephants Don't Play Chess." *Robotics and Autonomous Systems* 6, nos. 1–2: 3–15.

Brooks, Tim, Bill Peebles, Connor Holmes, et al. 2024. "Video Generation Models as World Simulators." OpenAI, February 15. https://openai.com/research/video-generation-models-as-world-simulators.

Burr, David, and Simon Laughlin. 2020. "Horace Barlow (1921–2020)." *Current Biology* 30, no. 16: R907–R910.

Cardon, Dominique, Jean-Philippe Cointet, Antoine Mazières, and Liz Carey-Libbrecht. 2018. "Neurons Spike Back." *Réseaux* 211, no. 5: 173–220.

Cassirer, Ernst. 1955. *The Philosophy of the Enlightenment*. Translated by Fritz C. A. Koelln and James P. Pettegrove. Beacon Press. Originally published in 1951 by Princeton University Press.

Cheselden, William. 1728. "An Account of Some Observations Made by a Young Gentleman, Who Was Born Blind, or Lost His Sight So Early, That He Had No Remembrance of Ever Having Seen, and Was Couch'd Between 13 and 14 Years of Age." *Philosophical Transactions of the Royal Society of London* 35, no. 402: 447–50.

Condillac, Étienne Bonnot de. 2014. "A Treatise on the Sensations." In *Philosophical Works of Étienne Bonnot, Abbé de Condillac*, vol. 1, translated by Franklin Philip. Taylor and Francis.

Descartes, René. 1998. *Discourse on Method and Meditations on First Philosophy*. 4th ed. Translated by Donald A. Cress. Hackett.

Diderot, Denis. 1749. *Lettre sur les aveugles à l'usage de ceux qui voient.* London. https://gallica.bnf.fr/ark:/12148/btv1b8613353x.

Diderot, Denis. 1772. *Lettre sur les aveugles à l'usage de ceux qui voient.* Amsterdam. https://wellcomecollection.org/works/rzy4d4gb.

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

Edelman, Shimon, and Lucia M. Vaina. 2001. "Marr, David (1945–80)." In *International Encyclopedia of the Social and Behavioral Sciences*, edited by Neil J. Smelser and Paul B. Baltes. Pergamon.

Freedberg, David. 2017. "From Absorption to Judgment: Empathy in Aesthetic Response." In *Empathy: Epistemic Problems and Cultural-Historical Perspectives of a Cross-Disciplinary Concept*, edited by Vanessa Lux and Sigrid Weigel. De Gruyter.

Gibson, Eleanor J. 1982. "Foreword." In *Reasons for Realism: Selected Essays of James J. Gibson*, edited by Edward Reed and Rebecca Jones. Lawrence Erlbaum Associates.

Gibson, James J. 1950. *The Perception of the Visual World*. Houghton Mifflin.

Gibson, James J. 1978. "The Ecological Approach to the Visual Perception of Pictures." *Leonardo* 11, no. 3: 227–35.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Gombrich, E. H. 1987. "Western Art and the Perception of Space." In *Space in European Art*. Council of Europe Exhibition, Tokyo.

Hamraie, Aimi, and Kelly Fritsch. 2019. "Crip Technoscience Manifesto." *Catalyst: Feminism, Theory, Technoscience* 5, no. 1: 1–33.

Hinton, Geoffrey E. 1990. "Connectionist Learning Procedures." In *Machine Learning: An Artificial Intelligence Approach*, vol. 3, edited by Yves Kodratoff and Ryszard S. Michalski. Morgan Kaufmann.

Irwin, Julia A. 2024. "Artificial Worlds and Perceptronic Objects: The CIA's Mid-Century Automatic Target Recognition." *Grey Room*, no. 97 (Fall): 6–35.

LeCun, Yann. 1988. "A Theoretical Framework for Back-Propagation." In *Proceedings of the 1988 Connectionist Models Summer School*, edited by David Touretzky, Geoffrey Hinton, and Terrence Sejnowski. Morgan Kaufmann.

LeCun, Yann. 2022. "A Path Towards Autonomous Machine Intelligence (Version 0.9.2)." June 27. https://openreview.net/pdf?id=BZ5a1r-kVsf.

LeCun, Yann. 2023. Interview with Craig Smith. *Eye on AI* podcast, episode 150 (November 2). https://www.youtube.com/watch?v=vjrU4MWjd6Y.

LeCun, Yann. 2024. Twitter/X post, February 19. https://x.com/ylecun/status/1759486703696318935.

LeCun, Yann, B. Boser, J. S. Denker, et al. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1, no. 4: 541–51.

Leffler, Christopher T., Stephen G. Schwartz, Eric Peterson, Natario L. Couser, and Abdul-Rahman Salman. 2021. "The First Cataract Surgeons in the British Isles." *American Journal of Ophthalmology* 230 (October): 75–122.

Malevé, Nicolas, and Katrina Sluis. 2023. "The Photographic Pipeline of Machine Vision; or, Machine Vision's Latent Photographic Theory." *Critical AI* 1, no. 1–2. https://doi.org/10.1215/2834703X-10734066.

Marr, David. 1980. "Visual Information Processing: The Structure and Creation of Visual Representations." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 290, no. 1038: 199–218.

Marr, David. 1982. *Vision*. W. H. Freeman.

Millière, Raphaël, and Cameron Buckner. 2024. "A Philosophical Introduction to Language Models, Part I: Continuity with Classic Debates." Preprint, arXiv, January 8, 2024, 2401.03910.

Mitchell, W. J. T. 1994. *Picture Theory: Essays on Verbal and Visual Representation.* University of Chicago Press.

Molyneux, William. 1996. "Letter to John Locke, July 7 [1688]." In *Molyneux's Problem: Three Centuries of Discussion on the Perception of Forms*, edited by Marjolein Degenaar. Springer.

Moravec, Hans. 1988. *Mind Children.* Harvard University Press.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83, no. 4 (October): 435–50.

Nelson, Randal C. 1991. "Vision as Intelligent Behavior: Research in Machine Vision at the University of Rochester." *International Journal of Computer Vision* 7, no. 1: 5–9.

Offert, Fabian. 2023. "On the Concept of History (in Foundation Models)." *image* 37, no. 1: 121–34.

Robinson, Howard M. 1982. *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rosenblatt, Frank. 1961. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Cornell Aeronautical Laboratory.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323, no. 6088: 533–36.

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton. 2021. "Reward Is Enough." *Artificial Intelligence* 299 (October). https://doi.org/10.1016/j.artint.2021.103535.

Sterne, Jonathan. 2021. *Diminished Faculties: A Political Phenomenology of Impairment.* Duke University Press.

Voltaire. 1738. *The Elements of Sir Isaac Newton's Philosophy.* Translated by John Hanna. Stephen Austen. Accessed via Gale Primary Sources. GALE|CW0106784756.

Werbos, Paul J. 1974. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." PhD diss., Harvard University.

Willems, Roel M. 2011. "Re-Appreciating the Why of Cognition: 35 Years After Marr and Poggio." *Frontiers in Psychology* 2 (September). https://doi.org/10.3389/fpsyg.2011.00244.

[3]

Vector Media

> How does Stable Diffusion work without an internet connection? [. . .] I always assumed it searched Google or something for ideas but then I read it doesn't access the internet. [. . .] Where is the "brain" that holds all that data? [. . .] Does each model hold all that info? The models are around 2–6 gb apiece, it really doesn't seem like they could hold that all.
>
> —Realistic7362, r/StableDiffusion, 2023

> This release 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.
>
> —Stable Diffusion public release, August 2022

ChatGPT Is *Not* a Blurry JPEG of the Web

"ChatGPT is a blurry JPEG of the web," wrote science fiction author Ted Chiang in early 2023 in *The New Yorker.* Chiang's metaphor is arresting and oddly precise: He is suggesting that ChatGPT, and large language models more broadly, do not reproduce the internet so much as compress it. They smooth over detail and blur structure; they return only a statistical approximation of meaning. "Hallucinations," for Chiang, are nothing more than compression artifacts.

Fundamentally, large language models are built on a kind of probabilistic text completion: on finding the most plausible set of next words for a set of given words. They occasionally produce completions that *sound* factual, but their truth status is entirely coincidental. The most prominent example is the hallucination of references to academic papers that do not exist—but with

plausible titles, journal names, authors, and so on. This, for Chiang, is simply the model interpolating from incomplete, "lossy" information. This interpolation gives the impression of comprehension: "When we're dealing with sequences of words, lossy compression looks smarter than lossless compression" (Chiang 2023).

Lossy compression, in ChatGPT, then, does not fill the traditional pragmatic function of compression. Or rather, its aim is different from the kinds of compression we have become used to dealing with: zip, mp3, mp4, jpeg, and so forth all aim to reconstruct a version, however distorted, of the original.

ChatGPT, for Chiang, is simply another form of lossy compression. Plagiarism masked by paraphrasing, such that the text produced by it seems new, even if it isn't. But why are we so impressed by its apparent intelligence, when nobody would do so for an mp4 file or a jpeg? Because, for Chiang, we read its "lossy" compression as lossless. And while "there's nothing magical or mystical about writing," he concludes, "it involves more than placing an existing document on an unreliable photocopier and pressing the Print button."

Chiang, a trained computer scientist, is able to trace ideologies in the specific properties of a technical object; he sees that a critique of AI must look into the black box. But here, unfortunately, Chiang's analogy is a rather unhelpful miss. Deep learning models (including ChatGPT) *are* compression systems, but of a fundamentally different kind—epistemologically and ideologically—from the sorts of compression systems we are used to. As we shall see in this chapter, this new form of *epistemic compression* extracts traditional digital media (such as image and text files) into infinitely commensurable vector embeddings, rendering machine vision systems into *vector media.*

Epistemic Compression

Computer scientists have traditionally distinguished between two fundamentally different kinds of compression: *lossy* and

lossless. Information that has been compressed losslessly can be reconstructed in full: If you put files from your computer into a zip archive, you can (at least in theory) always get back to the original versions. This is done by reducing redundancy at a statistical level, often by spotting identical sequences that reoccur and encoding them more efficiently. In visual terms: An all-white image could be encoded by saving the color value white as many times as the image has pixels, or it could be encoded by saving the *instruction* "repeat white as many times as the image has pixels."[1]

The aim of lossy compression, on the other hand, is to produce an approximation of the original signal by exploiting our own perceptual inattention to certain kinds of information—for instance, by eliminating frequencies too high to hear or see. Lossy compression algorithms are therefore more interesting ideologically as—by deciding what to keep and what to discard—they each contain a kind of immanent value theory of a particular medium, as Hito Steyerl (2009) has argued in her theory of the "poor image."

We ought here to pay tribute to perhaps the most important critical analysis of a single compression algorithm: Jonathan Sterne's (2012) work on the mp3 format. While ostensibly a media history of a file type, Sterne's book is, at heart, a political epistemology of compression. He argues that the mp3, as a form of lossy compression, "carries within it practical and philosophical understandings of what it means to communicate, what it means to listen or speak, how the mind's ear works, and what it means to make music" (2). The mp3 discards what it assumes no one will hear, and thus enshrines a model of an imagined human subject in decidedly imperfect listening conditions as the implicit end user of all recorded sound.

What is removed from the signal is not removed arbitrarily, but according to a psychoacoustic model, developed in military and industrial research settings, of how hearing is presumed to function. The mp3's particular genius thus lies in the invisibility (or inaudibility) of its losses—its ability to seem transparent, natural, faithful to the source, even as it discards large parts of it. This

compression creates what Sterne calls "perceptual capital"—or even, more accurately, "imperceptual capital, since it is generated from the accumulation of media material that users do not perceive" (48). This—the question of capital—is precisely what we will return to in the next chapter.

Sterne's critique remains indispensable in a media landscape still saturated with traditional compression formats. But if we are to seriously consider large language models and other neural networks, on some level, as compression systems, we must recognize that today's neural compression systems work rather differently.

We have seen that lossy compression formats succeed culturally when they appear to maintain fidelity to a presumed original, even as they silently enact value judgments about what deserves to survive. The logic of compression in artificial intelligence is not primarily based on perception, but on prediction; they have no psychometric model of a presumed human observer. They act at the epistemological level, rather than at the level of a particular medium; and they have little to do with reducing file sizes—indeed, they often increase them.

Neural networks, then, are algorithms of *epistemic* compression, in that they claim quite explicitly to model knowledge rather than merely reconstructing data of a presumed original. M. Beatrice Fazi, in her piece on "Synthesis in Generative AI," makes a somewhat similar observation by also understanding Chiang's analogy as deficient. Fazi (2024, 35), however, argues in particular for a better philosophical framing of the generative aspects of artificial intelligence systems, one that aims to reassess the "representational character of these technologies in a way that goes beyond mimetic assumptions." Fazi defines the synthetic as "a mode of thought but also a mode of production of and from thought. The synthetic, in this sense, is less what AI generates (its outputs) than the processes that it employs in this generation" (55). As we shall see, we can make an even stronger claim: Epistemic compression, in the age of artificial intelligence, is understood as a mechanism

not only for the abstraction of existing knowledge, but also for the production of *new* knowledge.

Claude Shannon, in his theory of information from which the "standard model" of compression (as the removal of redundancy) is commonly derived, had excluded questions of meaning, content, or semantics from his analysis of signal communication: "The fundamental problem of communication," he writes, "is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem" (Shannon 1948, 1). In a sharp reversal of this stance, today's epistemic compression systems are explicitly designed to hold information *beyond* what is necessary for reconstruction, and to explicitly model its "content." The blurriness of contemporary systems that Chiang observes, then, is not a technical fault, nor a move to "hide" incorrectness (as in Chiang's diagnosis). It is instead a by-product precisely of the search for an epistemic faculty in machine learning systems: a technical and ideological convergence in machine learning research, over the past forty years, of the concept of compression with that of *embedding*.

Compression Meets Embedding

An *embedding* (noun, the end product of the process of embedding data, verb) can be loosely defined, in our context, as a high-dimensional[2] numerical representation (or vector) of some kind of information unit: an image, a text, a sound, or perhaps some other media artifact. This conversion from traditional electronic media into a vector embedding is also known as a projection: For reasons which will become apparent, it is in effect a more general manifestation of forms of perspectival projection known to art historians.

Embedding is only one of a host of tightly correlated terms for this kind of work; the fact that much of the vocabulary is ill-defined

points to the contingent, and ultimately ideological, nature of embeddings themselves. It is a term often used in natural language processing, where one speaks of *word embeddings*, *sentence embeddings*, et cetera—though in the technical papers presenting some of the most important embedding techniques, such as word2vec and BERT, authors prefer to speak of "representations" (Mikolov et al. 2013; Devlin et al. 2019). In computer vision, which had a tradition of extracting "image features" through handwritten (not machine learning) algorithms such as SIFT, such spaces have also sometimes been known as *learned image features.* Finally, the term *latent space* has recently been gaining traction in the critical sphere, including through important ongoing work by Antonio Somaini (2025). Latent space describes the (vector) space in which embeddings sit; one speaks of "walking the latent space" through linear interpolation between embeddings. Often associated with generative adversarial networks (GANs), latent space is a term sometimes (though not exclusively) used in less formal contexts. To take an example—one to which we will return—in the paper presenting Alec Radford and colleagues' (2015) deep convolutional GAN, the emphasis is on "reusable feature representations"; latent spaces feature far more heavily in the GitHub README. We will, for now, stick to the term *embedding,* partly because it is slightly more general than latent space, and partly to circumvent the quagmire of representation, which we will discuss later in the chapter.

A key point of embeddings is the establishment of a *commensurability* of meaning. When processed by a neural network, two similar photographs of the same dog should result in two numerically similar vector embeddings, whereas comparing the embedding representations of one of those photographs to that of an image of, say, a microscope slide should result in a rather larger quantitative distance. Embeddings, in other words, are spaces of limitless comparison, sometimes, as we will discuss in the next chapter, even between media. As with all projections, embeddings encode these comparisons *from a particular point of view*: They are vector spaces with value systems. "A translation of psychophysiological

space into mathematical space," as Erwin Panofsky (1991, 66) put it in 1927; "in other words, an objectification of the subjective." He was famously describing perspectival transformations in art, but they are in fact closely related to embeddings.

How might one construct an embedding space? We might measure many different individual aspects of a signal. For a digital image, for instance, we could focus on its main colors, its contrast, its brightness, its—to return to Shannon—entropy. This approach, sometimes called the *handcrafted feature vector*, was an important technique of machine learning–based computer vision systems well into the 2010s. The feature vector would effectively be the image's virtual calling card. As a toy example, we might chain together the average hue, contrast, brightness, and entropy of each image to produce a 4D feature vector. Under the right circumstances, similar images ought then to have similar feature vectors, which could be compared mathematically. The feature vectors of past decades were of course more sophisticated than that—but even so, the kind of visual knowledge that we are able to model this way is disappointingly trivial. An image of a red vase might be calculated as similar to images of the same vase taken in very similar circumstances, but different from a blue vase, or even to the same red vase photographed in different lighting conditions or from a different angle.

Today, the algorithms doing the projecting are usually neural networks—not only because they are powerful algorithmic tools in general, but because by their nature they tend to contain embeddings at a structural level. Because conventional neural network architectures tend to proceed sequentially from input to output,[3] one can "cut" most neural networks at any layer, taking the activations—the internal representations of the neural network—at some intermediate point between the input (text, images, etc.) and output (perhaps classifications, or in some cases, more text, images, etc.). But we do not call just any layer of the network an embedding space: It is a space that claims to somehow represent the *contents* of the signal being encoded.

Gradually, computer scientists realized that neural networks in particular are very good at coming up with "meaningful" embedding spaces for images. From millions of examples, neural networks learn a way to represent an image, without being constrained to its low-level properties. In fact, they seem to produce them as almost an inevitable by-product of learning any task at all on an image dataset. This capacity seems to be inherent, for instance, in deep convolutional networks—the dominant form of computer vision algorithm over the past decade.[4] They work by chaining together a series of "layers," each of which is a relatively straightforward filtering operation called convolution.[5] A network is "deep" (as in "deep learning") if it combines many (tens or even hundreds of) layers. Each of these layers is, in general, in effect a mathematical representation of the layer preceding it, with the input image as layer zero. Each one of these layers is therefore some sort of representation (or embedding) of the layer preceding it, and thus of all previous layers, and ultimately of the input image.

The remarkable capacity of embeddings in deep convolutional networks can be seen, for instance, in neural style transfer. Much has been written about how poorly such systems implement something like an artistic style, or how problematic such a notion might be in the first place—but they become far more interesting if viewed as forensic visualizations, or infographics (Salvaggio 2023) of a neural network's embeddings. The original neural style transfer algorithm effectively works by attempting to compromise between the embeddings of two images within VGG-19 (Gatys et al. 2016). This is possible even though VGG-19 was not *explicitly* designed to produce embeddings, but rather to classify an image into one of one thousand discrete categories (which, of course, are the categories of ImageNet-1k). Instead, the embeddings of later layers, in being mathematically closer to the final layer (a prediction of output category), are highly correlated with the objects depicted in the image, whereas earlier layers are closer to low-level image statistics such as textures and colors. Neural style transfer thus worked by blending the earlier-stage embeddings (style) with

embeddings from a much later layer (content). Other applications would take the penultimate layer of VGG-19 (closer, though not identical, to Gatys's "content" layers) and use it for downstream tasks, such as visual similarity search or image quality estimation.

We ought, however, to distinguish between embeddings from VGG-19 that are epiphenomenal, and those that are, so to speak, learned explicitly: The latter is the domain of *representation learning.* Representation learning can sometimes work on explicit annotations, as in CLIP (which we will discuss in the next chapter), or it can operate without labels—"unsupervised learning" through, for instance, autoregression or reconstruction. In autoregressive systems, the model learns by predicting part of its input from another part: for example, learning word embeddings by predicting a word from its surrounding context, as in word2vec (which we will return to later in this chapter). In contrast, autoencoders compress the entire input into a lower-dimensional internal representation via an encoder, and then attempt to reconstruct the original input through a decoder. The central constraint is called a bottleneck: a narrow intermediate layer that forces the model to distill the input into its most essential features. What passes through the bottleneck is, in effect, a compressed code, an embedding that captures only what the model needs in order to rebuild the input.

Almost all neural network models, then, produce embeddings in one way or another, whether implicitly or explicitly. But does that mean they are compressing data, in the way Chiang imagines, or in any other meaningful sense? When, exactly, is embedding a form of compression? Compression, as a contemporary artificial intelligence paradigm, is not about resource management or the reduction of storage space. The gigabytes of space that contemporary models take up are orders of magnitude removed from the storage capacity of even off-the-shelf personal computers today. Random access memory size (especially on graphics processors) plays a role as well, given that models have often to be moved "to device" completely to be run efficiently—but even here it seems like only a matter of time until our personal computers, phones,

watches, and fridges are going to be running foundation models locally. Ironically, the engineering task of fitting a large neural network onto a mobile phone (so-called *model compression*) is much more like compression of the old form.

With a handful of well-publicized exceptions, even the largest language or image models will not be able to reproduce specific data exactly; there is simply not the space to do so reliably in a model that takes up far fewer gigabytes than the training data. In part, this is a matter of scale: The model's total parameter count is far smaller than the cumulative size of its training corpus, so it cannot store everything. But that limitation is not structural. Models can be trained to reproduce inputs with near-perfect fidelity, and in recent years, researchers have explored using language and multimodal models explicitly as compression engines, capable of lossless or near-lossless data recovery (Jamil et al. 2023). At the other extreme, we can imagine embeddings that do not compress at all. A very "wide" network applied to small images might produce representations that are larger (i.e., higher-dimensional) than the inputs themselves.

Neural embeddings, then, are a new form of compression—one that may or may not actually reduce dimensionality or file size in the traditional sense. What defines them is not efficiency, but epistemic function: Embedding becomes a means of producing new knowledge through the structuring of information into usable, translatable, extractable form. In contemporary AI, embedding *is* knowledge, and compression is the route to embedding.

Though epistemic compression has only recently become the dominant technical model in AI, we can trace its longer intellectual history to a concern with visual and cultural knowledge, in particular through the intersection of three independent tendencies—the first two of which we will explore in the rest of this chapter. Firstly, an attempt to build arithmetical and geometric regimes of representation—especially of visual-perceptual and cultural-linguistic data. In short, the cultural technique[6] of *embedding*.

Secondly, the development of the idea of compression, and especially *neural* compression—inspired especially by the optic nerve as a "bottleneck." And thirdly, an increasing attention to modeling information between media: toward media convergence and media indifference, as we will explore in the next chapter. Not merely a remediation (Bolter and Grusin 2000), but a true *media collapse*, in both a technical and media-theoretical sense, in which traditional media renounce their specificity as media. As we will discover, this media collapse implies a new techno-epistemic model founded only on commensurability and exchangeability, in which cultural artifacts are subsumed into *neural exchange value.* But to understand this fundamental—but largely invisible—paradigm change, we have to first disentangle the histories, and epistemologies, of the technologies at play.

Arithmetics of Meaning

Since the 1970s, gradually but surely, the ideological function of embeddings within computer science has shifted: from merely attempting to represent knowledge, to claiming to *produce* it. How is such a claim possible? It is at heart an epistemic claim of computer science, which we understand as the core theoretical and ideological intervention of contemporary artificial intelligence research. And it results from an *arithmetics of meaning.*

The mathematical theory underlying embedding in a more general sense is the theory of *vector spaces*, a branch of analytical geometry in mathematics, first formulated in this sense (Châtelet 1993; Fearnley-Sander 1979) by the nineteenth-century German mathematician Hermann Graßmann. Graßmann, later an important linguist,[7] was for most of his career a schoolteacher whose contributions to mathematics were underacknowledged because of the overly philosophical tone of his work. He was fond, for instance, of Friedrich Schleiermacher's ideas about dialectics (Petsche 2009) and opens his book, titled *Ausdehnungslehre*, with a quite thorough reflection on the epistemology of the natural sciences (thoughts

from things) compared to the epistemology of mathematics (thoughts from thoughts).

Graßmann's crucial move was to abandon the metaphors of three-dimensional space—to show that there is nothing special about three-dimensionality, and thus to abstract space from physical geometry. He called this theory "extension theory" (*Ausdehnungslehre*), which was to be "the abstract basis of a theory of space (geometry) [. . .] stripped of all geometric content" (Graßmann 1878, 277). For Graßmann, extension theory was to geometry as algebra is to numbers: a way to generalize spatial-geometric relations.

Graßmann's writings were underappreciated at the time (with a few exceptions, such as Hermann von Helmholtz), but rose to prominence through the influential Italian mathematician Giuseppe Peano, known for the axioms underlying set theory. "The words segment, area, volume ought to be substituted by vector, bivector, trivector," writes Peano (2000, xiv), emphasizing Graßmann's move from three-dimensional geometric space (the kind that Immanuel Kant was able to argue was *synthetic a priori*, i.e., emerged from spatial intuition) to multidimensional abstract geometry. The idea of a high-dimensional vector space was to become incredibly important in the natural sciences in the decades following Graßmann's and Peano's theoretizations, as well as—eventually—in computer science (Arianrhod 2024).

In an 1853 paper, Graßmann applied his notion of vector space to the problem of color. Thomas Young and Helmholtz had already shown that human retinas are sensitive to three types of color (very roughly red, green, blue)—but Graßmann is the first to propose such a thing as the *color space*, which is a vector space of color, such that "points close to each other in the color signal space are perceived also subjectively close, that is, they produce a small color difference between them" (Logvinenko 2015, 4). An abstract geometric space, in other words, in which visual-perceptual similarity is matched to similarity in the color space. We come back once again to Panofsky on perspective (1991, 66):

a "translation of psychophysiological space into mathematical space; in other words, an objectification of the subjective." Graßmann had here anticipated both the mathematics and the epistemics of the embedding space.

The color space is at its heart a representation, not a compression. Its aim is not to reconstruct the original. And although any set of numerical descriptions of an object might be considered a vector embedding, the notion of an embedding space as a model of culture comes about more explicitly in the late 1960s and early 1970s, in Gerald Salton's work on information retrieval through the SMART information retrieval system at Cornell (Salton et al. 1975).[8] SMART is now better known as TF-IDF (term frequency–inverse document frequency), a way of modeling information where each dimension of the vector space tells you how often a particular word is used. This representation throws information away, such as the ordering of words, but it is not compressed in the sense of a meaningful (even lossy) recoverability of the original information. It is an essentially destructive model of documents, where each point in a high-dimensional space corresponds to one document, and where similar documents appear close by. That vector space might be a useful tool to analyze anagraphic (i.e., fundamentally discrete) data such as text, then, is a much later addition than its use in geometry and color science—but one that lays the foundations for the later epistemic turn in artificial intelligence.

Salton's TF-IDF is an example of what we described above as a handcrafted embedding: there is no complex statistical learning involved in the technique. But over a decade after Salton's original work, a technique called latent semantic analysis (Landauer et al. 1998) applied a simple and well-known statistical compression algorithm, principal component analysis (PCA), to the TF-IDF matrix. The resulting vector representation is not directly a compression of the text, but a compression of an already highly restricted embedding representation. Embedding and compression have met—though we shall encounter several other, historically unrelated, cases of this collision.

Nevertheless, the combination of compression and embedding present in latent semantic analysis led to a radical shift in how people conceptualized the algorithm: Compression forces a reorganization of the embeddings, effectively translating discrete symbols into continuous signals. TF-IDF (embedding without compression) had been talked about essentially as a library search algorithm. Latent semantic analysis—precisely the same technique, but with compression—was then seen as something capable not only of *pointing to* meaning, but of *encoding and even creating it* by interpolating between points, in a manner not dissimilar from exploring the latent space of a contemporary AI model. Rather than merely a search algorithm, latent semantic analysis was presented explicitly as a "biologically and psychologically plausible mechanistic explanation of the acquisition, induction, and representation of verbal meaning" (Landauer 1999, 303), noting that "dimension-reduced vector and matrix representations [. . .] have long been regarded as serious models of psychological process" (306). Cognitive theories based essentially on the metaphor of compressed linguistic embeddings have continued in the decades after Landauer, before the recent explosion of interest into large language models (Gärdenfors 2014), and most certainly since. Yet, as we shall see, the intellectual history of epistemic compression is bound up not primarily with language, but with vision.

The Roots of Epistemic Compression

David Marr, the central psychologist of vision we met in the last chapter, credits his former colleague at Cambridge Horace Barlow with being the first person to connect the psychology of vision quite directly to neural activity. Barlow's radical claim was that neural activity *alone* could account for the nervous system. Marr (1982, 13–14) tells of his enthusiasm for the "vigor and excitement" of Barlow's account, leading to his doctoral work on the cerebellar cortex: "Truth, I also believed, was basically neural." In a passage quoted by Marr, Barlow writes of the significance of his own work: "Each *single neuron can perform a much more complex and subtle task than had previously been thought*" (emphasis is Marr's).

Barlow's work is especially important in the development of computational thinking around vision not just because it manages to connect neural processes with perceptual-psychological ones ("the activities of neurons, quite simply, are thought processes"; Barlow 1972, 380, cited also in Marr 1982), but because of the hierarchical and compressional nature of the neural visual pathways. Much of the emphasis of the neuroscience of vision was thus shifted from the visual cortex to the retina. In a 1961 article for *Sensory Communication*, Barlow posits three possible principles around which visual neurons may be organized (though, as he admits, they may not be mutually exclusive). Firstly, the password hypothesis: Specific stimuli might be encoded uniquely, such that the neural representation of a visual form might be an arbitrary "code." Secondly, as "control points" that modulate the flow of information downstream. Or third—Barlow's own model—that biological neural networks might be essentially compression machines; that the "reduction of redundancy is an important principle guiding the organization of sensory messages and is carried out at relays in the sensory pathways" (Barlow 1961, 217). He thus emphasizes that the primary task of sensory relays is not merely to pass on signals, but to encode them in a way that minimizes unnecessary data, while retaining critical new information. In this sense, Barlow's vision of the brain's operations connects the physical structure of neural pathways to an overarching strategy of optimizing information flow—what he calls the "economy of impulses."[9]

Barlow then develops his claim that the nervous system maximizes the efficiency of information transmission by discarding redundant or repeated sensory inputs in the mathematical terms of Shannon's information theory. His perspective through the 1950s was undoubtedly shaped by Shannonianism and cybernetics, including through his engagement with the Ratio Club, a cybernetics dining club that also counted among its members Donald Mackay, Alan Turing, and Giles Brindley, David Marr's PhD supervisor. But where Shannon's theory focused on the efficiency of message transmission in noisy channels, famously indifferent to the semantic content of the signal, Barlow introduces a kind of physiological

attention: Not only do neural architectures focus on the "meaningful" part of perceptual signals, but different organisms are likely to process different sorts of phenomena as important. It was Barlow's attention to the optic nerve that fully crystallized his compression model. Barlow's view of the relatively simple *retinal* neurons as key agents in visual processing is crucial here; as the optic nerve is a kind of bottleneck between the eye and the brain, it follows that the retinal neurons must learn to do something like compression of visual data. In Barlow's (1969) account, then, sensory processing becomes an active "economy of impulses," optimizing information flow by amplifying the novel and unexpected while discarding the predictable.

By 1958, Barlow had a wider and more ambitious version of this claim, which he presented at the now-famous (British) National Physical Laboratory's conference on the Mechanisation of Thought Processes. Though vision remains his key case study, he argues that sensory mechanisms across the brain operate like compression algorithms—extracting essential information while filtering out what is predictable or irrelevant. He proposes that this process of reducing redundancy underlies the efficiency of sensory processing and is a defining feature of intelligent systems, whether biological or artificial. This aligns with the ideas of O. G. Selfridge's *Pandemonium* paper from the same conference, which describes visual processing in hierarchical stages, a perspective that became foundational for computational models of vision. Compression, in other words, is the first step to intelligence, as Barlow (1969, 210) was later to phrase it: "Furthermore, the operations required to find a less redundant code have a rather fascinating similarity to the task of answering an intelligence test, finding an appropriate scientific concept, or other exercises in the use of inductive reasoning. Thus, redundancy reduction may lead one towards understanding something about the organization of memory[10] and intelligence, as well as pattern recognition and discrimination." It would be wrong to think of Barlow's perspective on intelligence as unchallenged in the history of artificial intelligence. The expert systems or GOFAI (good

old-fashioned AI, i.e., based on explicit symbolic rules), popular well into the first decade of this century, were based on something like the opposite of compression: They branched out endlessly into an (ultimately abandoned) effort to capture the breadth of human thought. These systems aimed to map out an exhaustive representation of human knowledge, leading to immense knowledge bases that were designed to house explicit, structured rules for every imaginable scenario. One such project in continuous development since 1984, Cyc, contains at least twenty-five million "microtheories" about everything from geopolitics to botany, each an attempt to capture a microcosm of knowledge about the world: closer to what Antonio Gramsci called *senso comune* (something like "common sense"), the misleadingly unstable bedrock of ideology.

Neocognitrons

Another key participant at the 1958 National Physical Laboratory conference was Frank Rosenblatt. It is at this point in most standard histories of neural networks and computer vision that the story shifts decisively to the invention of the perceptron. And rightly so: Rosenblatt's machine, publicly demonstrated in July of that year, was the first statistically trainable neural network implemented in hardware: a breakthrough often narrated as the genesis of neural networks.

As James Dobson (2023) recounts in detail, Rosenblatt left the Cornell Aeronautical Laboratory shortly after the 1958 announcement and took up a faculty position at Cornell University, where he continued to refine and expand his models. The later stages of his work complicate the standard narrative: Dobson is clear to dispel the myth that Rosenblatt simply built a shallow network and moved on. At Cornell, he constructed the Tobermory, an audio-classification machine with four layers and over twelve thousand adjustable weights—far more complex than the single-layer version targeted in the famous critique by Minsky and Papert.

Rosenblatt also sought to correct the misreadings of his work that had emerged in the wake of its media reception. In the 1961 volume *Principles of Neurodynamics*, he opened with a pointed preface, distancing his project from the inflated claims of artificial intelligence and reaffirming its status as a model of biological intelligence. "A perceptron," he wrote, "is first and foremost a brain model, not an invention for pattern recognition" (Rosenblatt 1961, vii). The insightful account by Dominique Cardon (2018) emphasizes this point as well: that Rosenblatt should not be seen as a failed proto-AI engineer, but as a theorist committed to neurodynamic explanations of intelligence.

In the final years of his life, Rosenblatt was conducting speculative experiments in biological information transfer, while also becoming involved with the Air War Study Group at Cornell, which worked to expose the human, political, and ecological costs of the Vietnam War.[11] As James Dobson notes, this marked a striking reversal: the inventor of systems designed for military photointerpretation turning to critique the very conflicts such systems had helped make operational.

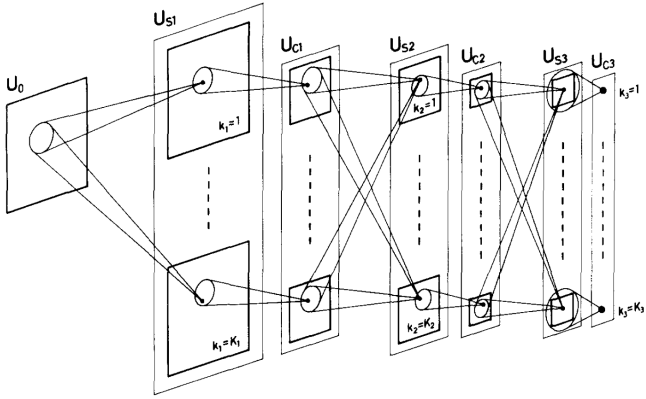
Building on the work of Barlow, who had been among the first to measure neural activity for vision directly, David Hubel and Torsten Wiesel engaged in a series of breakthrough experiments throughout the 1960s. As with the Molyneux experiment and the Cheselden case discussed in the previous chapter, Hubel and Wiesel's work was inspired by blindness (and especially the question of why removing congenital cataracts—i.e., those present since birth—should not immediately restore sight to a patient); their experiments thus famously involved suturing shut the eyes of kittens and recording their neural responses with invasive electrodes. Their work identified two primary types of cells within the cats' visual cortex: so-called simple cells, which respond to specific visual features (e.g., edges and orientations); and complex cells, which seem to integrate information from groups of simple cells. Thus they were able to articulate more precisely the basic model Barlow had already suggested, and which was to inspire subse-

quent computational models through Marr's dogma in particular: a *hierarchical* model of visual information processing. Hubel and Wiesel shared the 1981 Nobel Prize in Medicine for "discoveries concerning information processing in the visual system."

Hubel and Wiesel's work directly inspired the Japanese engineer Kunihiko Fukushima, working at NHK, the Japanese public broadcaster's research lab, in the 1970s. Fukushima was tasked with exploring compression techniques for high-definition TV signals. NHK was interested in sending higher definition signals using the existing American NTSC analog television format Japan had adopted (525 lines at 30 frames per second); rather than working on this compression algorithm directly, Fukushima was tasked with working on a biomimetic system which would approximate and quantify the perceptible loss in image quality of the compressed TV signal.[12] His background in signal compression led him to consider how visual perception itself could be modeled through hierarchical networks that reduce data complexity. Fukushima's (1980) invention, the *neocognitron*, was an attempt at a simulation of the simple and complex cells identified by Hubel and Wiesel: a multilayered neural network designed to perform pattern recognition by filtering and combining visual features in a hierarchical manner.

The neocognitron's architecture features two primary types of artificial neurons, which Fukushima called S-cells and C-cells (respectively models of simple and complex cells). S-cells are responsible for extracting local features from the input, while C-cells allow the network to tolerate small positional shifts in these features. This configuration enabled the neocognitron to perform simple forms of pattern recognition, including the recognition of Japanese characters and Arabic numerals. In a very real sense, it was the first example of a convolutional neural network of the kind used today for computer vision, which successively passes filters on top of an image signal: a network quite unlike the "artificial neural networks" of Rosenblatt.

Fukushima's convolutional network was intended to be a "self-organized" model of vision, "learning without a teacher" based on



[Figure 3.1]. Diagram of the neocognitron from Fukushima (1980). Courtesy of the author.

the "gestalt" of the input images: an explicit example of learning without reinforcement or reward, or "unsupervised learning" (Fukushima 1980, 193). He later (Fukushima 2007) describes the neocognitron's self-organizing dynamics as learning a "division of labor." Unsupervised learning is one of the paradigms that emerges naturally from Barlow's compression model of cognition. In an influential[13] paper on the topic in 1989, Barlow asks, "What use can the brain make of the massive flow of sensory information that occurs without any associated rewards or punishments?" (Barlow 1989, 295).

The neocognitron may have been "self-organized," through a rather clunky reinforcement-update mechanism; but the mechanics of training a neural network were still poorly understood by 1980. The question of selecting appropriate visual stimuli to detect in S-cells (or equivalently, of learning convolutional *filters*) is a highly complex mathematical problem, which could be attempted only once one had an algorithm for how to change the weights of each neuron in response to some kind of training stimulus or dataset. This was the backpropagation algorithm, popularized[14] within AI research by

D. E. Rumelhart, G. E. Hinton, and R. J. Williams in 1986 (Rumelhart et al. 1986, 533); it allows for "an arbitrarily connected neural network to develop an internal structure that is appropriate for a particular task domain." Backpropagation remains, in some form, the way we train almost all contemporary neural networks; yet the algorithm is presented not for its superior ability to solve tasks, but to generate *hidden* (latent) representations while learning these tasks: "Internal 'hidden' units which are not part of the input or output come to represent important features of the task domain." Yann LeCun,[15] who had also worked on refining the backpropagation algorithm (LeCun et al. 1988), applied it to the convolutional network invented by Fukushima: The result was LeNet (LeCun et al. 1989), a convolutional network trained to decode handwritten zip codes. The MNIST image dataset of digits—handwritten by U.S. Census Bureau employees and high school students—was developed by the U.S. National Institute of Standards and Technology as part of LeCun's project, and is widely used to this day as a classic dataset for teaching computer vision techniques.

The neocognitron and its successors, while computational models of vision, are better understood as conceptual machines rather than mere programs. They embody not just an architecture for processing visual input but also a theory of media—specifically, the medium of television. In Fukushima's work, the network is both an analogy for how visual signals might be compressed and reconstructed in human perception and a framework for handling the immense complexity of high-definition visual data. It models television signals through hierarchical processes of signal decomposition and recomposition, where layers of processing filter, condense, and amplify the most significant aspects of an image. When presenting the neocognitron, Fukushima (1980, 193) emphasizes that the "network has been simulated on a digital computer"—the network itself is a formal thought experiment, and any actual computational work is a mere simulation of it. LeCun, too, speaks of training LeNet as "simulations [. . .] performed using the backpropagation simulator SN [. . .] running on a SUN-4/260."

(LeCun et al. 1989, 546). These models are inspired by biological processing, but they are not formal neuroscientific models: Scholars at the time are clear that their designs "emphasize computational power rather than biological fidelity" (Hinton 1990, 556). These early convolutional networks, then, are caught between conceptual model and pragmatic algorithm: They both model a specific medium (television) and universal perception (vision), serving simultaneously as models of compressed media and universal models of compression itself. They are simultaneously tied to the material and cultural specificity of their medium and striving toward an abstract, medium-agnostic ideal. This striving for sublimation, as it will turn out, returns in contemporary models with a vengeance.

The Analogy Machine

By the late 1980s, Horace Barlow had already established redundancy as the central problem of perception: not a nuisance to be avoided, but the very material from which meaning is extracted. In his 1989 paper on unsupervised learning, he gave this idea its most forceful formulation: Redundancy, he argued, is not merely what vision removes, but the very condition of perception itself. "The technically correct term, redundancy, is almost equally misleading, for it suggests that this part of the sensory inflow is useless or irrelevant, whereas it is the potential source of all the available knowledge about the constant or semiconstant patterns and regularities in an animal's environment. *Knowledge* is perhaps the best term for it, though it may seem paradoxical that this knowledge of the world around us can be derived only from the redundancy of the messages" (Barlow 1989, 297–98).

Without repetition—without the patterns, constraints, and regularities that define the environment—there would be nothing to perceive, nothing to learn. It is only through statistical saturation that the brain can construct any model of the world at all.

This idea found practical traction in the 1990s, when computational neuroscience began to treat natural image statistics not as noise,

but as the empirical ground of visual coding. A key demonstration came from Bruno Olshausen and David Field (1996), who—citing Barlow—trained an unsupervised model to reconstruct image patches using a minimal number of active units ("sparse coding"), replicating the presumed energy minimization of biological neural networks. Though biologically motivated, their model was essentially computational—and yet it learned filters that closely resembled the receptive fields of neurons in the primary visual cortex. Olshausen and Field's influential finding[16] was more than a technical breakthrough; it embodied the new ideology of intelligent compression.

The theoretical work of computational neuroscientists in the 1990s was matched by developments in the use of compression as a pragmatic approach to training neural networks. In the 1980s, Erkki Oja (1982) showed that simple networks could approximate principal component analysis,[17] the prototypical algorithm of statistical compression, to which we will later return. A few years later, Pierre Baldi and Kurt Hornik (1989) demonstrated that multilayer networks could extend this to what they called "non-linear PCA," leading to the now-familiar autoencoder architecture that we introduced above: an input compressed through a central bottleneck and then reconstructed. The bottleneck was not just a constraint but a principle. By forcing the network to distill its input into fewer dimensions, autoencoders formalized a view of learning as compression. Though less biologically grounded and mathematically elegant than sparse coding, autoencoders were arguably even more influential. Their basic architectural pattern—encoder, bottleneck, decoder–quickly became foundational, and continues to underpin much of generative AI today, both directly (autoencoders are present in transformer- and diffusion-based image models) and through related encoder-decoder architectures such as U-Net.

Compressed embeddings, then, have existed for some decades in machine learning. They gradually rose in prominence, but became a quasi-hegemonic technical paradigm perhaps only in the 2010s. A cornerstone work here is Tomas Mikolov and colleagues' 2013 paper "Efficient Estimation of Word Representations in Vector

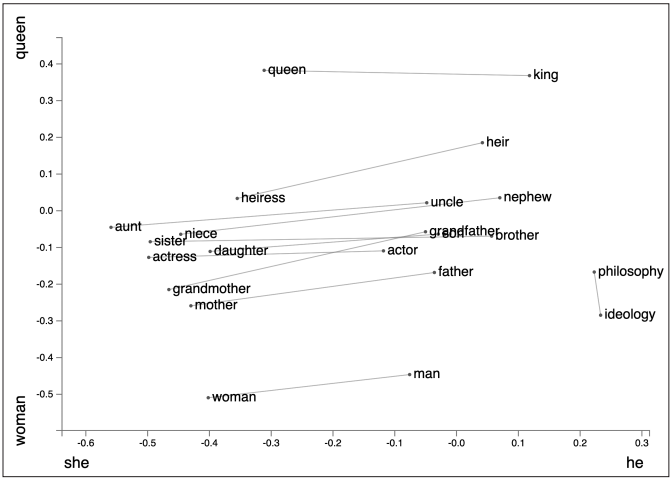
Space," which first mentioned the fact that embedding models could be evaluated through "vector representations, with the expectation that not only will similar words tend to be close to each other, but that words can have multiple degrees of similarity." Just a few years earlier, Turney (2006, 1) had already suggested that what he calls "relational similarity" might turn out to be an interesting measure of capabilities beyond syntax of vector spaces. It was also Turney who first proposed that the trope most amenable to the operationalization of relational similarity might be the *analogy*—which would go on to be one of the core technical metaphors of the embedding space.[18] It was Mikolov and colleagues' "skip gram" model, however, that popularized the idea that unsupervised language models could pick up such relational similarities from raw text data.

The textbook example of such an analogy in computer science (recalling the bizarre gender politics of Turing's Imitation Game, Rosenblatt's perceptron,[19] and much of artificial intelligence research since then) is the analogy "Man is to king, as woman is to what?" The answer—queen—is easily inferred, of course, but until relatively recently this was an extremely hard problem to solve computationally. The particular example is first mentioned in another paper by Mikolov, with Wen-tau Yih, and Geoffrey Zweig, titled "Linguistic Regularities in Continuous Space Word Representations" (2013, 5). What Mikolov and colleagues present, in the context of this analogy, is a so-called word-embedding model. Not unlike image embeddings, word embeddings capture—and compress[20]—a very large data corpus, such that meaningful relationships between words are preserved. The model proposed by Mikolov achieves this through a relatively "shallow" neural network architecture that is trained to predict a context window of about two to five words to the left and right from a "missing" word in the center of that context window. The middle (hidden) layer of the architecture consists of a fixed amount of neurons (in the original paper, between fifty and six hundred)—this layer becomes the embedding in the fully trained model. More concretely, when the

model is trained, the hidden layer is tuned such that it will always "route" the input signal in such a way that the output corresponds to the most likely solution to the "fill in the blank" task presented to it.[21]

Where, in this model, can we find the embedding "space"? It is important to note here—and this is maybe nowhere as obvious as in word-embedding models—that any embedding model just learns a projection. There is no "space" here to look at, just a mechanism to produce specific outputs from specific inputs. One of the most prominent misconceptions of humanities critiques of artificial intelligence relates to this somewhat counterintuitive fact. There is nothing "in" an embedding model except complex rules of transformation. This is not to say that these rules of transformation are meaningless, just math, or culturally insignificant. In fact, this book argues precisely *for* the cultural significance of embeddings. But we cannot simply remove bad pictures or words from a fully trained model without fundamentally changing its structure, including the internal logic of its similarity relations.[22] As Geoff Hinton (1986, 1), a key figure in the development of the notion of embedding—whose perspective we will treat in detail in the next chapter—puts it, "When the weights in the network are changed to incorporate new knowledge about one concept, the changes affect the knowledge associated with other concepts."

Mikolov's word-embedding model learns to associate a set of floating-point numbers—one number for each neuron in the hidden layer of the model (so between fifty and six hundred)—with every word in the training vocabulary. As with vision models, we can now simply disregard the output layer of the model (the part that gives us the predicted context window for the input word) and use the model as an encoder, from vocabulary word to set of numbers. And it is here where the "spatial" interpretation of embeddings begins to make sense again: We can now understand this set of n values as one point in n-dimensional space, which in turn allows us to put words into relation with each other, in the context of the trained word-embedding model. We can measure,



[Figure 3.2]. Arithmetics of meaning in vector space. The *y*-axis is essentially "power," the *x*-axis is "gender." Created using "Explore Word Analogies," n.d., https://lamyiowce.github.io/word2viz/.

for instance, the distance between two encoded words through simple geometric operations, as the arithmetic of high-dimensional vector spaces allows for almost the same kinds of operations that are possible in Euclidean space. Some caveats apply here that we will return to.

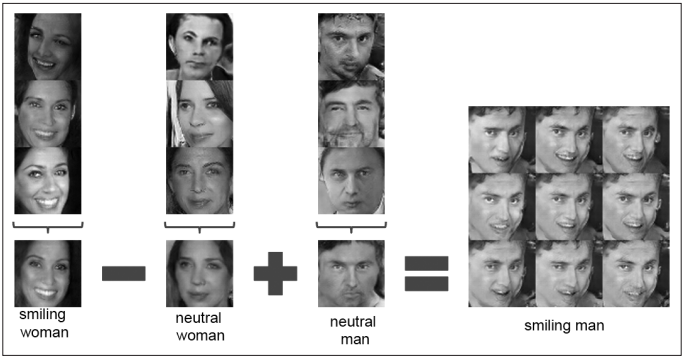
What Mikolov showed, then, is that distance within this projected embedding space is meaningful. Related words are geometrically close to each other, and unrelated words are geometrically distant. This makes intuitive sense, of course, as the model has learned to predict a context window, and meaningfully similar words will likely appear in similar contexts: a fundamental principle of natural language processing called the distributional hypothesis.[23]

The surprising aspect of this model is that this kind of spatial mapping is meaningful beyond the similarity of words. In fact—as hinted at earlier—it can be used to solve relational analogies. Concretely, if we compute the points for "king," "man," and

"woman," with the help of a word-embedding model similar to that of Mikolov and colleagues, and if we then complete the high-dimensional parallelogram that is suggested by the arrangement of these three points in vector space, the point that is closest to the ideal completion of that parallelogram will be the vector encoding of "queen": Man is to king as woman is to queen. What is perhaps even more surprising is that this capability of the model translates to a more general spatial semantics. The axis on which we can move from man to woman, and thus also from king to queen (this is why the geometric expression of the analogy is a parallelogram) equates, more generally, to gender. On it, we can also move from mother to father, or from aunt to uncle. Problematically—and this is exactly where the contemporary discourse on bias in artificial intelligence begins (Bolukbasi et al. 2016)—we can move, on the same axis, also from nurse to doctor, without moving on another axis (let's say "power") at the same time. This other axis does exist as well in the embedding space of the model, as it forms the "vertical" (in Euclidean) term shift of the parallelogram; it is the axis on which we move from man to king, and from woman to queen.

Radford and colleagues were the first to demonstrate, already in 2015, that a similar arithmetic of meaning can be implemented in the visual domain as well. Subsequent research, all the way to contemporary multimodal generative models, has time and again tried to disentangle axes of visual meaning. In fact, it could be argued that GANs were so popular not only because they were able to produce "realistic" images, but because their embedding space seemed to be so meaningful initially, with axes almost naturally aligned with conceptual changes.

To put this contemporary research into historical perspective, clearly all attempts to operationalize away meaning—with or without contempt for the "soft" scholarship so attached to it—can be traced almost all the way back to Leibniz's (1989) first attempts at a universal language that would not be riddled with problems of ambiguity and untranslatability. In fact, the development is almost circular, given that recent research into large language models has



[Figure 3.3]. Arithmetics of meaning in the visual domain. Figure from Radford et al. (2015).

proposed to utilize them for the production of political consensus (Tessler et al. 2024)—exactly what Leibniz[24] intended with his *universal characteristic*, which was meant to resolve conflict before it could even arise. What is there to discuss if we can simply compute? We should, in other words, view the many attempts by computer science to demonstrate the meaningfulness of their vector spaces with some skepticism.[25] Not because we would like to argue for a fundamentally unquantifiable, unoperationalizable "core" of the human experience—this is not necessary at all to arrive at a valid criticism of vector arithmetics as an arithmetics of meaning. Instead, let us introduce a few clarifications that, together, already point us to more fundamental limitations of the method, and, eventually, to some significant political implications. All these clarifications have to do—perhaps not surprisingly—with the fact that the arithmetics of meaning, demonstrated so convincingly by Mikolov and colleagues, takes place within high-dimensional vector space.

The first clarification pertains to the idea that embedding models can solve analogy questions. In reality, the solution to the analogy query is given by the model not as a definite answer, but as a hierarchy of existing data points. Why? Simply because there are no intermediate words. If the best possible analogy is a (new) data point right in between two (existing) data points representing

words in the source corpus vocabulary, the best possible analogy is neither of them, but still can only be described in terms of them. Even if the input vocabulary consisted of all words in the English language, the vector operations that solve the analogy query could still produce a data point between all points.[26]

Every computational solution to an analogy task is thus, ironically, itself an analogy—one that, in a peculiar reversal of the usual narrative of violent quantification so often adopted by critiques of computation, is a violent "qualification" of the machine's solution. We are "bending" the precise solution offered by the model to fit our discrete model of language. Our interpretation of any results produced by a word-embedding model is therefore, in fact, an interpretation of an interpretation. In other words, much of the meaningfulness of computational solutions to analogy questions is that we are willing to overlook the interpolation necessary to return to the discrete space of words from the quasi-continuous embedding space. We assume that the solution to the analogy query is the hierarchy of references to existing high-dimensional data points presented to us by the algorithm, while, in fact, this hierarchy is a *visualization* of the underlying vector space solution.

Why a "visualization"? High-dimensional vector spaces—this is the second clarification—are geometrically counterintuitive spaces. Not only is it impossible to imagine a vector in eleven dimensions, high-dimensional vector spaces have counterintuitive properties as well—this is what is known as the "curse of dimensionality." One of the core problems of higher dimensions is that "under certain broad conditions [. . .], as dimensionality increases, the distance to the nearest neighbor approaches the distance to the farthest neighbor" (Beyer et al. 1999, 217). In other words, in high-dimensional vector space, distances between data points have a tendency to become illegible: They lose, or at least significantly change, their meaning. Thus they become inaccessible to intuition. More generally, for us as humans, the only geometrically intuitive space is Euclidean space (n = 3). We simply do not have the mental capabilities to think spatially about anything above n = 3.

Consequently, to be able to think spatially about a larger space, we have to find a way to map the larger space to Euclidean space, or, in other words, visualize the larger space.

This means—and this is the third clarification—when a model produces a legible (i.e., *visible*) answer to an analogy query, it actually produces two visualizations: a numerical visualization and a geometric visualization. The numerical visualization is the hierarchy of references to existing high-dimensional data points presented to us by the algorithm. The geometric visualization is the "collapsed" high-dimensional space, presented to us as the computed real (i.e., floating point) values of these data points.

Operationalizing Representation

We have seen in detail how word2vec and similar embedding models do the work of epistemic compression. But the underlying technique of embedding is actually more prominent, to the point where we could say that contemporary artificial intelligence, to a large part, relies on embedding. In fact, we could look at almost any contemporary model and, at least for some part of its pipeline, identify an encoder-decoder architecture that effectively implements an embedding layer. We could almost speak of an iconography of epistemic compression, a deeply geometric—in both its actual functioning and its representations—way to understand knowledge, a bottleneck epistemology.

To return to Jonathan Sterne (2012, 250), we are not interested in replacing "a grand narrative" of learning "with a grand narrative of ever-increasing compression," but we are indeed "proposing compression as one possible basis for inquiry into the history of communication technology—in the same sense that representation has served." In other words, we are interested here in the epistemology of a concrete computational technique—embedding—but the history of this technique, if understood broadly as a cultural technique, is much longer, and deeply intertwined with the question of representation in general.

In fact, one could understand embedding as the logical next step in the historical debate on pictorial representation discussed in the previous chapter. There, the main question turned out to be whether pictures are reality—that is, are two-dimensional representations of a three-dimensional world sufficient to learn a meaningful model of that world, a world model, in computer science terminology? Now, the main question becomes: What is the epistemic status of embeddings, the result of another (but much stranger) transformation? Are embeddings to pictures what pictures are to reality?

This line of questioning is part of our attempt to reevaluate the overloading of philosophical terms by computer science. Computer science constantly claims to "do" representational work, to produce "good," "meaningful," or "disentangled" representations. Previous critiques have often assumed that the use of this and similar terms in the technical disciplines is necessarily reductive—that is, terms like "representation" are used in a narrow technical sense, bracketing its cultural and philosophical implications. With Phil Agre's understanding of the technical roadblocks in computer science *as* effects of philosophical challenges, we would like to instead take seriously computer science's claim to operationalize—technically *and* epistemically—what it means to represent reality.

Crucially, the process of embedding includes two epistemically relevant transformations: First, we "add continuity" to discrete data—words become floating-point vectors—only to take it away again when we want to make the "new knowledge" visible that emerges this way. We should keep in mind here that the first step of "expansion" does not fully apply to images, which are already "more continuous" than text. We will return later to this important factor. All this is relevant not only to embeddings, but has a deeper, more fundamental connection to the process of representation in general. Indeed, the basic problem of geometric perspective is exactly one of *dimensionality reduction*: How do you represent a three-dimensional scene in a two-dimensional image?

This translation between three-dimensional space and two-dimensional space is called, by both art historians and computer scientists, *projection.*

Shadows are also forms of projections. In one of the origin myths of painting, Kora of Sicyon traced the outline of her lover's shadow on a wall. The projection, in analytical geometry, into lower-dimensional space has sometimes been called the *shadow*—for instance, perhaps not surprisingly, by Hermann Graßmann, the first person to conceive of vector space as a tool for analytical geometry—who calls it *Abschattung* (shadow).

Let us restate our claim more succinctly: *Dimensionality reduction is an operationalization of representation itself*, just as embedding is the operational form of meaning. The two are inseparable, and indeed often overlapping. Together, they model a form of sense-making that does not follow the linearity of logical reasoning, but instead operates through quasi-geometric regularities: spatial relations that imply, rather than directly denote, meaning.

As M. Beatrice Fazi (2025, 146) convincingly argues, "Computing machines are representational machines"—not in the sense that they reflect reality, but in that they generate what she calls an "internal representational reality" (140). To understand the epistemic force of this representational machinery, Fazi insists that "we need a model of synthesis that can address—boldly and directly—this representational character" (146). Such a model would take seriously the capacity of artificial systems to construct, not merely infer or represent, phenomena: to produce meaning through structure, rather than to decode it from signal.

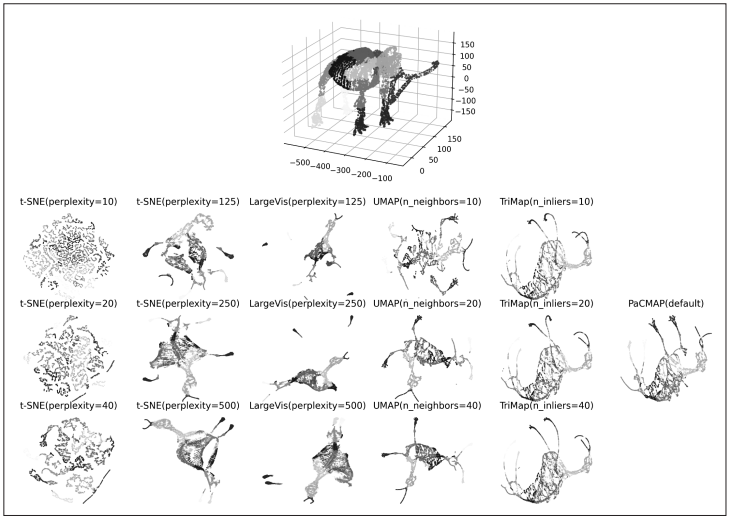
This is why the exclusive understanding of artificial intelligence as "pattern recognition" misses the mark. Or rather, it underestimates the kind of pattern recognition at stake. These are not patterns of the image space—edges, contours, bounding boxes[27]—but patterns within an abstract embedding space. To return to the vector arithmetic that underlies the king-queen analogy, the parallelogram that "solves" this analogy exists in a space of such high

dimensionality that it resists visualization. We have to be aware that we will never, therefore, see the "full picture" of this argument. What we see instead is—perhaps not surprisingly at this point—a *lossy compression.*

To be clear, lossy compression is a useful technique, as any simple audio compression algorithm that removes inaudible frequencies outside the human perceptual range shows. Here, nothing of value is lost—but only because we know in advance which frequencies are audible and which frequencies we can safely eliminate. This is not the case, at all, for embeddings of cultural data if they are produced by a machine learning system. There, we can never be sure, exactly, what role a specific property of the probability distribution we attempt to model plays. In fact, one of the biggest—and to this day unreached—goals of machine learning research is to produce perfectly disentangled representations—that is, embeddings where we know which axes (which frequency, in signal processing terms) correspond to which meaningful aspect of the real world.

Take the case of dimensionality reduction, a process (and set of associated techniques) used to render high-dimensional data visible in Euclidean space. Just like all other forms of lossy compression, dimensionality-reduction algorithms try to preserve as much of the structure of the high-dimensional data as possible, such that, for instance, the relative distances of data points are preserved. Points far apart in high-dimensional space should remain far apart in Euclidean space. But as we are, again, dealing with lossy compression here, some information must be thrown away for this to work. Essentially, then, the task of the dimensionality-reduction algorithm is to decide which dimensions to keep and which to discard.

In the figure below, for instance, multiple forms of dimensionality reduction are used on a set of three-dimensional points, which together form the shape of a mammoth. On top, we see a parallel geometric projection, common in technical drawing, which is *itself* a dimensionality-reduction technique (you are looking at a two-dimensional image!). The other images of the mammoth are



[Figure 3.4]. Dimensionality reduction of a mammoth. Reproduced from Wang et al. (2021).

achieved through t-SNE[28] and UMAP, two common contemporary dimensionality-reduction techniques.

The t-SNE and UMAP plots still preserve some "meaning" of the original, three-dimensional point cloud. The points that make up specific body parts of the mammoth, for instance, are still clustered together. And yet, the most important "semantic" property of the three-dimensional data—that all points form the shape of a mammoth—is lost, and cannot be reconstructed at all from the plots. This argument—by extension—holds not just for projection that happens to be into two dimensions, but for all forms of projection, and thus for dimensionality reduction in general. There is, in other words, already a strong ideological component to be found in dimensionality reduction.

Art historians and media theorists will be familiar with the nature of this problem. That any projection is always an ideological choice is, as we noted earlier, already at the basis of Panofsky's *Perspective*

as Symbolic Form (1991, 66): "Exact perspectival construction is a systematic abstraction. [. . .] The result was a translation of psychophysiological space into mathematical space; in other words, an objectification of the subjective." This is later mirrored in Jean-Louis Baudry and Allan Williams's (1974) theory of the ideological function of the cinematic apparatus (now known as "apparatus theory"), or in Henri Lefebvre's (1991) analysis of the production of space. All three (would) agree: There is no such thing as an *independent* mathematical space; instead, we always need to take the entanglement of mathematical space, representational space, and lived space into account. Whenever we move from one to the other, decisions are being made, and wherever decisions are being made, ideologies come into play.

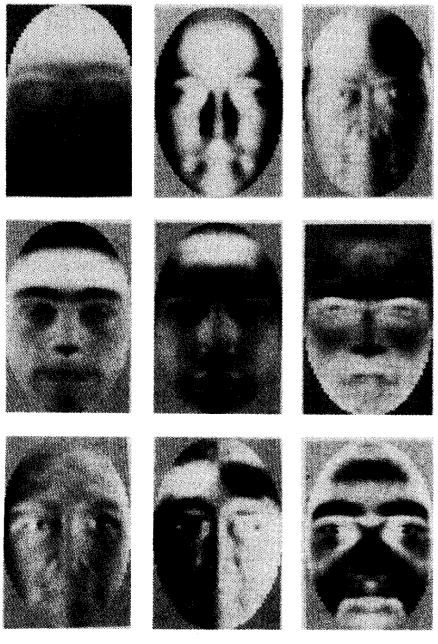
From its very conception, then, the idea of a vector-embedding space is linked to the idea of the quantification and rationalization of the *visual*, first in geometry and then in color science. The eventual—technical and ideological—multimodality of embeddings shines through already in its earliest historical manifestations, not only the technique itself, but also in those techniques necessary to make it work. Both (perspective and color spaces) are still integral to modern visual computing. And, in a sense, Graßmann's ideas about color spaces foreshadow the fundamentally multimodal character of embeddings—as a way to operationalize *away* media specificity; a theme we have mentioned throughout this book—and will return to in the subsequent chapter. We could summon, of course, the whole history of art to give evidence to this point—a book that will need to be written another time. Instead, we should attempt to summarize our argument so far. What becomes clear at this point is that all the core concepts we discussed are not only fundamentally entangled, but they are all significantly overloaded as well.

Representation in the technical sense means embedding, embedding is compression. At the same time, compression requires dimensionality reduction to be made legible, which is not only a generalization of representation in the philosophical sense but also an ideological choice. Truly in the spirit of Phil Agre, we are facing,

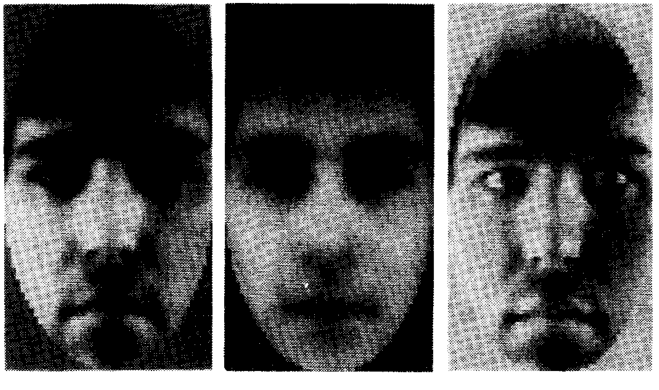
with the epistemic paradigm shift in artificial intelligence research, a "stack" of conjoined technical, philosophical, and sociopolitical problems that cannot simply be reduced to—as it were—either perspective. We can already see here the complexity that emerges from comparably simple models when we pay attention not only to their "semantic" capabilities, but to the ways in which these capabilities are tied to the concrete, mathematical, and technical architecture of the model. Most important, however, is the fact that we have limited our analysis, so far, to the desirable properties of embedding space—to the fact that embedding spaces can model at least some of the more advanced ways that humans have at their disposal to deal with "meaning." But embedding models are not limited to established ways of producing meaning. In fact, much stranger relations—simultaneously "meaningful" for the model and "meaningless" for us—must exist in embedding space.

It is unsurprising, then, that contemporary critical work on machine learning has pointed out, time and again, the ideological component of statistical techniques (Chun 2021) like dimensionality reduction, and most prominently perhaps has exposed the deep entanglement of statistical concepts and scientific racism in facial recognition, which can be traced all the way back to Karl Pearson and the PCA technique, to Francis Galton and composite photography, and to Charles Spearman and IQ testing.[29]

We will thus return one last time to facial recognition. There are many ways to implement facial recognition, but arguably the most prominent way has been a technique called eigenfaces (Turk and Pentland 1991; Lee-Morrison 2019), invented several years before the neural network resurgence of the early 2010s. Eigenfaces—a play on eigenvectors, which withstand linear transformations and thus contain some fundamental aspect of the phenomenon they describe—surprisingly use only a simple dimensionality-reduction algorithm, PCA, to achieve a very high success rate in recognizing faces. Eigenfaces are often successful—exactly because of their relation to eigenvectors—if faces appear "in the wild"—that is, they appear transformed due to being recorded from a different



[Figure 3.5]. Eigenfaces, reproduced from Michael Kirby and Lawrence Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, no. 1 (1990): 103–8. Reprinted with permission from IEEE.



perspective than the training samples the model has seen. This robustness is an effect of applying PCA to a large training dataset of faces. Any image dataset—as the reader should remember from the previous sections—can be understood as a collection of points in a space that has the size of the images in the dataset (including three color planes used to express RGB values). As with any other high-dimensional space, we can run a dimensionality-reduction algorithm on this data to preserve only the most meaningful dimensions—that is, those dimensions where the variance is greatest, and which thus encode most of the dataset. We can then—this is the trick of the Eigenfaces approach—express any new faces as a linear combination of these main components. Comparing images as such combinations, then, gives us a good-enough approximation of the similarity of two faces, even if they are photographed differently. To get a clearer picture of the inner workings of this process, we can visualize these meaningful dimensions—and we will notice that only some face components are legible to us.

The majority of eigenfaces, however, are illegible—or at least counterintuitive—for us. The embedding space of contemporary models is "filled" with what can only be described as "machinic" concepts (here: faces) and conceptual relations that are resurrected from the ruins of their human equivalents. Embedding space, in other words, becomes a trove of potential "alternative" ways of thinking about text and images alike. Here lies the root of the epistemic turn in artificial intelligence. Taking a closer look at embedding, then, seems worthwhile not only from the technical, but precisely from the critical perspective.

Compression Ideology

We began this chapter with Ted Chiang's provocation: What if ChatGPT is a blurry jpeg of the web? Compression is indeed central to the epistemology of models like ChatGPT—but not in the way that jpeg or zip files are compressed. Large language models do not merely shrink data for efficiency. They restructure it. They embed it into high-dimensional spaces from which new, plausible

forms can be generated. These are not photocopies; they are synthetic projections.

Nonetheless, computer scientists have investigated compression as a serious theoretical model for intelligence, as well as a practical way to generate useful embedding spaces. Gregory Chaitin, a theoretical computer scientist from IBM, gave a paper at the German Philosophical Society in 2002 with the ambitious title "On the Intelligibility of the Universe and the Notions of Simplicity, Complexity, and Irreducibility." Looping through Plato, Leibniz, Albert Einstein, and Voltaire, he argues for a kind of extreme Kolmogorovianism in which "comprehension is compression, i.e., explaining many facts using few theoretical assumptions, and that a theory may be viewed as a computer program for calculating observations" (Chaitin 2002, 1). Comprehension is, in other words, in this view, the discovery of the shortest generative rule for observed data.

Mathematician Matt Mahoney (2009) made the claim even more bluntly: "Text compression is equivalent to AI." In 2006, he cofounded the Hutter Prize, offering significant financial rewards for algorithms able to efficiently compress a fixed 1 GB extract of Wikipedia. Its aim, to be clear, was never archival efficiency, but rather a poor man's Turing Test. The prize was designed by Marcus Hutter, a key figure in theoretical AI, to formalize the link between compression and intelligence: to compress well is to predict well, and to predict well is to model the world. Under this logic, compression is not merely a computational operation, but a cognitive ideal, capable of benchmarking intelligence itself.[30]

In truth, the relationship between compression and intelligence is stranger still. We cannot simply say that models like ChatGPT are glorified zip files. But perhaps the reverse is true: zip files themselves[31] can, in some circumstances, behave like language models.

Take gzip, one of the most widely used file compression tools in the world. Released around the time of the web (1992), it has become a core part of internet infrastructure: Nearly every web page you visit is first compressed with gzip before being sent to your browser.

Its purpose is simple: make files smaller so they load faster and take up less space. But the way it does this is surprisingly elegant. If a word, phrase, or sequence appears more than once, gzip replaces later instances with a kind of pointer back to the first one, saving space instead of storing the same thing over and over (a process based on what's known as LZ77 sliding-window compression). Then it rewrites the whole file using shorter codes for the most common elements, so that frequent characters or symbols take up fewer bits than rare ones (a method called Huffman coding)—a little like acronyms or abbreviations. It doesn't interpret or paraphrase the content; it simply looks for repetition and builds a shorthand around that.

In 2023, researchers showed that the compressed form of text that naturally emerges from gzip makes for a representation that can be used as input to a simple classifier (Jiang 2023). Especially on smaller datasets, as a representation, gzip beat many traditional textual neural networks (e.g., the LSTM) and more recent transformer models (including BERT), and far outperformed them on out-of-distribution datasets (e.g., low-resource languages). gzip is mindless, shallow, and reversible—it had never been seen as an "intelligent" representation of the information. It was compression *without* embedding, precisely as we had described jpeg at the start of this chapter. Now, in Jiang's work, it became "intelligent" by use-case alone.

gzip shows how epistemic compression, through its ideological ambitions, has quite pragmatic consequences. For gzip (or any other algorithm) to be considered intelligent representation, an *embedding*, is a change in affordances, not ontology: The same technical object becomes an epistemic object depending on what you do, or claim you might do, with it. More precisely, as we shall argue in the next chapter, by virtue of its use for the production of *neural exchange value.*

Notes

1 The general-mathematical form of this kind of redundancy reduction, or compression, is formalized as Kolmogorov complexity: The description length of the most efficient computer program (minus constants like the length of names given to functions) to reproduce a given string is understood as its complexity (Kolmogorov 1965).

2 We use this term rather loosely, in that with embedding spaces we are almost always talking about more than three dimensions, and therefore spaces in which our spatial intuition is rather limited; though we note that within the technical literature a space of several hundred dimensions (say, the bottleneck in an encoder-decoder architecture) might be contextually referred to as "low-dimensional," in that it is still commonly lower in dimensionality than the input space.

3 Not quite all neural network architectures are precisely sequential in this way: There are small deviations, such as skip connections, and large deviations, such as Hopfield networks. But even in such cases, one can find neural activations.

4 The transformer architecture complicates this picture, as we will discuss in the next chapter.

5 We are simplifying here, of course; there are in practice many examples of nonconvolutional layers and nonsequential connections.

6 The term cultural technique refers to one of the fundamental concepts of German media theory that emerged from the Berlin and Weimar schools of media studies in the late 1990s and early 2000s. Its most concise definition comes from Thomas Macho (2003, 179), who understands it as a set of practices that come before their theoretization: "Cultural techniques—such as writing, reading, painting, counting, making music—are always older than the concepts that are generated from them. People wrote long before they conceptualized writing or alphabets; millennia passed before pictures and statues gave rise to the concept of the image; and until today, people sing or make music without knowing anything about tones or musical notation systems. Counting, too, is older than the notion of numbers. To be sure, most cultures counted or performed certain mathematical operations; but they did not necessarily derive from this a concept of number." See also Siegert (2023).

7 Graßmann, notably, studied theology, languages, and literature, but not mathematics.

8 For a glimpse into the complex history of this paper, see Dubin (2004); for a historical overview of information retrieval and the move "from frequencies to vectors" see Rieder (2020).

9 Barlow is not quite the first to make this connection; he cites, for instance, Attneave (1954).

10 There would be much to say about the slippage between memory as a neurophysiological metaphor and memory qua hardware, opening many interesting questions about temporality and the neural network, which we hope to address elsewhere.

11 Matteo Pasquinelli (2023) traces the perceptron's origins to mid-twentieth-century psychometrics, where statistical techniques originally designed to quantify human mental traits such as intelligence were repurposed for machine vision, and to its emergence within cybernetic paradigms that positioned pattern recognition as a control mechanism. As he argues, "Rosenblatt invented the first artificial neural network (the perceptron), he automated the technique of psychometrics for image recognition" (Pasquinelli in Anderson 2025).

12 Convolutional networks, of the sort pioneered by Fukushima, are still used for this task, including for evaluating other neural networks. This is now called a "perceptual loss function."

13 From Barlow's obituary in *Current Biology*: "One of the first to draw attention to the importance of unsupervised learning as opposed to supervised or reinforced learning. . . . Horace's information theory-based approach underlies many modern approaches to unsupervised learning in neural networks and Bayesian learning" (Burr and Laughlin 2020, 909–10).

14 Though invented prior to that, possibly independently, for instance by Paul Werbos in the 1970s.

15 Compression is never far away: LeCun (along with Yoshua Bengio) was subsequently heavily involved in image compression and the DjVu format.

16 See Simoncelli and Olshausen (2001) for a wider account of its influence.

17 Oja's proof is valid for networks with a single hidden layer, and is far more interesting than the "degenerate" proof that infinitely large neural networks can approximate arbitrary continuous functions under certain conditions, the so-called universal approximation theorem.

18 More specifically, an analogy of the form "A:B::C:D, meaning A is to B as C is to D." For instance, "traffic:street::water:riverbed. Traffic flows over a street; water flows over a riverbed. A street carries traffic; a riverbed carries water. There is a high degree of relational similarity between the word pair traffic:street and the word pair water:riverbed. In fact, this analogy is the basis of several mathematical theories of traffic flow."

19 It is no coincidence that contemporary research on bias starts with gender. Artificial intelligence, especially in the visual domain, has been preoccupied with gender as an "easy" test case for vision. Already in early research on perceptron-like machines in the 1950s and 1960s, distinguishing "men" from "women" (which is clearly always also a reference to the Turing Test) was presented as a demo, with the rock band the Beatles—and a wig-wearing judge—as edge cases.

20 Note that word embeddings do not necessarily "compress" an individual word in this sense, as the embedding for a word may well take up more disk space, so to speak, than the original sequence of letters. Rather, they can be seen as systems which compress their training dataset.

21 As with all machine learning models, the idea here is, of course, to then use the model on unseen data, in the hopes that it has "generalized"—that is, picked up something about the general transition probabilities informing natural language and not just the transition probabilities inherent in the training data.

22 On a superficial level, one can, of course, simply censor the vocabulary of a model (by, for instance, removing a derogatory term), and it will no longer be produced by a nearest-neighbors search—but changing the underlying epistemic structure remains close to impossible. See Offert and Phan (2025); Meng et al. (2022).

23 First discussed as "meaning at the collocational level" in Firth (1957), echoing Ludwig Wittgenstein's famous dictum that *meaning is use* (and especially use within a language game).

24 One of the few Enlightenment philosophers to answer the Molyneux question, described in the last chapter, in the affirmative.

25 On the question of the meaning of meaning in artificial intelligence research, see also Bajohr (2023) and Bunz (2019).

26 Crucially, this is not true for image models (e.g., GANs), which, again, has contributed to their popularity: There are, in fact, meaningful intermediate images to be found between any two images.

27 These were, however, often the focus of earlier and less successful paradigms of computer vision systems.

28 Both t-SNE, the t-distributed stochastic neighbor embedding technique and its "nondistributed" predecessor—this will not surprise the reader returning here from the subsequent chapter—were codeveloped by Geoff Hinton (Hinton and Roweis 2002; van der Maaten and Hinton 2008).

29 "Proposals that stimuli be modeled by points in a space in such a way that perceived similarity is represented by spatial proximity go back to the suggestions of Isaac Newton. [. . .] However, little progress was made toward the development of data-analytic methods for the construction of such spatial representations on the basis of psychological data until the efforts of a group of psychometricians, beginning in the late 1930s at Chicago and subsequently moving to Princeton, culminated in the 1952 development by Torgerson of the first fully workable method of metric multidimensional scaling" (Shepard 1980, 390).

30 We seem to have come full circle against Pasquinelli's (2023) reading of AI as a continuation of human intelligence metrics.

31 To be precise, tar.gz rather than zip files are encoded with gzip.

Bibliography

Anderson, Shane. 2025. "What Is AI?" Interview with Matteo Pasquinelli. *032c*, January 8. https://032c.com/magazine/what-is-ai-matteo-pasquinelli.

Arianrhod, Robyn. 2024. *Vector. A Surprising Story of Space, Time, and Mathematical Transformation.* University of Chicago Press.

Attneave, Fred. 1954. "Some Informational Aspects of Visual Perception." *Psychological Review* 61, no. 3: 183–93.

Bajohr, Hannes. 2023. "Dumb Meaning: Machine Learning and Artificial Semantics." *image* 37, no. 1: 58–70.

Baldi, Pierre, and Kurt Hornik. 1989. "Neural Networks and Principal Component

Analysis: Learning from Examples Without Local Minima." *Neural Networks* 2, no. 1: 53–58.

Barlow, Horace B. 1961. "Possible Principles Underlying the Transformation of Sensory Messages." In *Sensory Communication*, edited by Walter A. Rosenblith. MIT Press.

Barlow, Horace B. 1969. "Trigger Features, Adaptation, and Economy of Impulses." In *Information Processing in the Nervous System: Proceedings of a Symposium Held at the State University of New York at Buffalo, 21st–24th October, 1968*, edited by K. N. Leibovic. Springer.

Barlow, Horace B. 1972. "Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?" *Perception* 1, no. 4: 371–94.

Barlow, Horace B. 1989. "Unsupervised Learning." *Neural Computation* 1, no. 3: 295–311.

Baudry, Jean-Louis, and Alan Williams. 1974. "Ideological Effects of the Basic Cinematographic Apparatus." *Film Quarterly* 28, no. 2: 39–47.

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent. 2000. "A Neural Probabilistic Language Model." In *Advances in Neural Information Processing Systems* 13 (2000): 932–38.

Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. "When Is 'Nearest Neighbor' Meaningful?" In *Proceedings of the 7th International Conference on Database Theory (ICDT '99)*, edited by Serge Abiteboul and Peter Buneman. Springer.

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

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *Advances in Neural Information Processing Systems* 29: 4349–57.

Bunz, Mercedes. 2019. "The Calculation of Meaning: On the Misunderstanding of New Artificial Intelligence as Culture." *Culture, Theory and Critique* 60, nos. 3–4: 264–78.

Burr, David, and Simon Laughlin. 2020. "Horace Barlow (1921–2020)." *Current Biology* 30, no. 16: R907–R910.

Cardon, Dominique, Jean-Philippe Cointet, Antoine Mazières, and Liz Carey-Libbrecht. 2018. "Neurons Spike Back." *Réseaux* 211, no. 5: 173–220.

Chaitin, Gregory J. 2002. "On the Intelligibility of the Universe and the Notions of Simplicity, Complexity and Irreducibility." Preprint, arXiv, October 2, 2002, math/0210035.

Châtelet, Gilles. 1993. *Figuring Space: Philosophy, Mathematics, and Physics.* Kluwer Academic Publishers.

Chiang, Ted. 2023. "ChatGPT Is a Blurry JPEG of the Web." *New Yorker*, February 9. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Chun, Wendy Hui Kyong. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition.* MIT Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding."

In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics.

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

Dubin, David. 2004. "The Most Influential Paper Gerard Salton Never Wrote." *Library Trends* 52, no. 4: 748–64.

Fazi, M. Beatrice. 2024. "The Computational Search for Unity: Synthesis in Generative AI." *Journal of Continental Philosophy* 5, no. 1: 31–56.

Fazi, M. Beatrice. 2025. "A Transcendental Philosophy of Large Language Models." *Philosophy and Digitality* 2, no. 1: 133–49.

Fearnley-Sander, Desmond. 1979. "Hermann Grassmann and the Creation of Linear Algebra." *American Mathematical Monthly* 86, no. 10 (1979): 809–17.

Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930–1955." In *Studies in Linguistic Analysis*, edited by John R. Firth. Blackwell.

Fukushima, Kunihiko. 1980. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36, no. 4: 193–202.

Fukushima, Kunihiko. 2007. "Neocognitron." *Scholarpedia* 2, no. 1. https://www.doi.org/10.4249/scholarpedia.1717.

Gärdenfors, Peter. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces.* MIT Press.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. 2016. "Image Style Transfer Using Convolutional Neural Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Graßmann, Hermann. 1853. "Zur Theorie der Farbenmischung." *Annalen der Physik* 165, no. 5: 69–84.

Graßmann, Hermann. 1878. *Die lineare Ausdehnungslehre: Ein neuer Zweig der Mathematik, dargestellt und durch Anwendungen auf die übrigen Zweige der Mathematik wie auch auf die Statik, Mechanik, die Lehre vom Magnetismus und die Krystallonomie erläutert.* Otto Wigand.

Hinton, Geoffrey E. 1986. "Learning Distributed Representations of Concepts." In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, edited by Jerry R. Hobbs and Robert C. Moore. Lawrence Erlbaum Associates.

Hinton, Geoffrey E. 1990. "Connectionist Learning Procedures." In *Machine Learning: An Artificial Intelligence Approach*, vol. 3, edited by Yves Kodratoff and Ryszard S. Michalski. Morgan Kaufmann.

Hinton, Geoffrey E., and Sam Roweis. 2002. "Stochastic Neighbor Embedding." In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, edited by Suzanna Becker, Sebastian Thrun, and Klaus Obermayer. MIT Press.

Impett, Leonardo, and Fabian Offert. 2024. "There Is a Digital Art History." *Visual Resources* 38, no. 2: 89–108.

Jamil, Sonain, Md. Jalil Piran, MujibUr Rahman, and Oh-Jin Kwon. 2023. "Learning-Driven Lossy Image Compression: A Comprehensive Survey." *Engineering*

Applications of Artificial Intelligence 123 (August). https://doi.org/10.1016/j.engappai.2023.10636.

Jiang, Zhiying, Tianyu Gao, Xinyu Zhang, Danqi Chen, and Jason Eisner. 2023. "'Low-Resource' Text Classification: A Parameter-Free Classification Method with Compressors." In *Findings of the Association for Computational Linguistics, ACL 2023.* Association for Computational Linguistics.

Kirby, Michael, and Lawrence Sirovich. 1990. "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, no. 1: 103–8.

Kolmogorov, Andrei N. 1965. "Three Approaches to the Quantitative Definition of Information." *Problems of Information Transmission* 1, no. 1: 1–7.

Landauer, Thomas K. 1999. "Latent Semantic Analysis: A Theory of the Psychology of Language and Mind." *Discourse Processes* 27, no. 3: 303–10.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25, no. 2–3: 259–84.

LeCun, Yann, David Touretzky, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1988. "A Theoretical Framework for Back-Propagation." In *Proceedings of the 1988 Connectionist Models Summer School*, vol. 1, edited by David Touretzky. Morgan Kaufmann.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1, no. 4: 541–51.

Lee-Morrison, Lila. 2019. *Portraits of Automated Facial Recognition: On Machinic Ways of Seeing the Face.* Transcript Verlag.

Lefebvre, Henri. 1991. *The Production of Space.* Translated by Donald Nicholson. Blackwell.

Leibniz, Gottfried W. 1989. "Preface to a Universal Characteristic" and "Samples of the Numerical Characteristic." In *Philosophical Essays*, translated by Roger Ariew and Daniel Garber. Hackett.

Logvinenko, Alexander D. 2015. "The Geometric Structure of Color." *Journal of Vision* 15, no. 1: 1–9.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9 (November): 2579–605.

Macho, Thomas. 2003. "Zeit und Zahl: Kalender und Zeitrechnung als Kulturtechniken." In *Bild–Schrift–Zahl*, edited by Sybille Krämer and Horst Bredekamp. Wilhelm Fink.

Mahoney, Matt. 2009. "Rationale for a Large Text Compression Benchmark." Last updated July 23, 2009. https://mattmahoney.net/dc/rationale.html.

Marr, David. 1982. *Vision*. W. H. Freeman.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. "Locating and Editing Factual Associations in GPT." *Advances in Neural Information Processing Systems* 35: 17359–72.

Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." Preprint, arXiv, last revised September 7, 2013, 1301.3781.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Offert, Fabian, and Thao Phan. 2024. "A Sign That Spells: Machinic Concepts and the Racial Politics of Generative AI." *Journal of Digital Social Research* 6, no. 4: 49–59.

Oja, Erkki. 1982. "Simplified Neuron Model as a Principal Component Analyzer." *Journal of Mathematical Biology* 15, no. 3: 267–73.

Olshausen, Bruno A., and David J. Field. 1996. "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images." *Nature* 381: 607–9.

Panofsky, Erwin. 1991. *Perspective as Symbolic Form.* Zone Books.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence.* Verso.

Peano, Giuseppe. 2000. *Geometric Calculus According to the Ausdehnungslehre of H. Grassmann.* Translated by Lloyd C. Kannenberg. Birkhäuser.

Petsche, Hans-Joachim. 2009. *Hermann Graßmann.* Birkhäuser.

Radford, Alec, Luke Metz, and Soumith Chintala. 2015. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." Preprint, arXiv, November 19, 2015, 1511.06434.

Rieder, Bernhard. 2020. *Engines of Order: A Mechanology of Algorithmic Techniques.* Amsterdam University Press.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323, no. 6088: 533–36.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18, no. 11: 613–20.

Salvaggio, Eryk. 2023. "How to Read an AI Image: Toward a Media Studies Methodology for the Analysis of Synthetic Images." *image* 19, no. 1: 83–99.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (July, October): 379–423, 623–56.

Shepard, Roger N. 1980. "Multidimensional Scaling, Tree-Fitting, and Clustering." *Science* 210, no. 4468: 390–98.

Siegert, Bernhard. 2023. *Kulturtechniken: Rastern, Filtern, Zählen und andere Artikulationen des Realen.* Rombach.

Simoncelli, Eero P., and Bruno A. Olshausen. 2001. "Natural Image Statistics and Neural Representation." *Annual Review of Neuroscience* 24, no. 1: 1193–216.

Somaini, Antonio (editor). 2025. *Le Monde selon l'IA: Explorer les espaces latents*. JBE Books.

Sterne, Jonathan. 2012. *MP3: The Meaning of a Format.* Duke University Press.

Steyerl, Hito. 2009. "In Defense of the Poor Image." *e-flux* 10, no. 11. https://www.e-flux.com/journal/10/61362/in-defense-of-the-poor-image.

Tessler, Michael Henry, Michiel A. Bakker, Daniel Jarrett, et al. 2024. "AI Can Help Humans Find Common Ground in Democratic Deliberation." *Science* 386, no. 6719: 796–801.

Turk, Matthew A., and Alex P. Pentland. 1991. "Face Recognition Using Eigenfaces."

In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society.

Turney, Peter D. 2006. "Similarity of Semantic Relations." *Computational Linguistics* 32, no. 3: 379–416.

Wang, Yingfan, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization." *Journal of Machine Learning Research* 22, no. 201: 1–73.

[4]

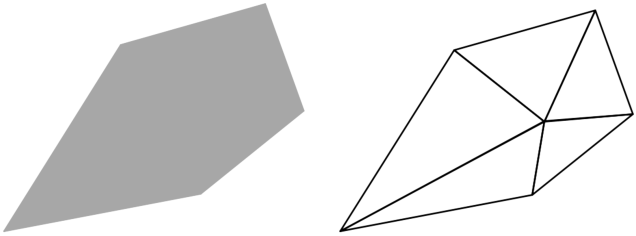
Neural Exchange Value

> In *Capital*, Marx puts forward the idea that there must be some common underlying essence shared by all goods. [. . .] The same philosophical point seems to apply to hypotheses [about digital images].
>
> —Geoffrey E. Hinton, *Relaxation and Its Role in Vision*

The Geometry of Value

Consider the two digital images—or rather, the two shapes—below (Figure 4.1). They might have been produced by some kind of computer graphics software; they are made up entirely of straight edges, and as such are called *polygons*. How might a simple computer algorithm compare the two? We first need to define a common, commensurable unit. We might choose their center of mass or their perimeter, but the simplest common measurement would seem to be the area of the shapes. We then need a way of calculating these areas. Though we might know the formulas for squares, rectangles, regular pentagons, and so forth, it would be impractical to include every such formula in our computer program. In any case, there are more exotic shapes for which simple formulas do not exist.

A better solution, then, is to split our shapes up into triangles, measure the area of each triangle, and add them up. This process will be familiar to practically any student of computer science as *polygon triangulation*. Any virtual 3D object in video games or rendered computer graphics is made up of such triangles. But this process of polygon triangulation, as a way of *collapsing different kinds of image into the same metric space*, is by some coincidence



[Figure 4.1]. Polygon triangulation.

the metaphor[1] Karl Marx (1976, 127) chooses to introduce the concept of exchange value: "In order to calculate and compare the areas of rectilinear figures, we decompose them into triangles. But the area of the triangle itself is expressed by something totally different from its visible figure, namely, by half the product of the base multiplied by the altitude. In the same way the exchange values of commodities must be capable of being expressed in terms of something common to them all, of which thing they represent a greater or less quantity."

Exchange value is perhaps the central concept in Marx's theory of the commodity. As a theory, Marx is keen to point out, it is far wider and more general than simply an account of currency or money. Money, indeed, is simply the exchange form that has happened to dominate in nineteenth-century industrial Western society, and is itself highly historically heterogeneous as a form: for instance, fiat versus reserve currency, pegged versus independent, metal-based versus paper. Rather, exchange value is first and foremost a quantification, a commensurability rather than an equivalence, always "characterized by a total abstraction" (Marx 1976, 127)—a much more violent concept than the term suggests. Price, then, is just one instance of exchange value, expressed in his time in terms of money. But any regime of exchange value will, for Marx, tend naturally toward the development of a universal equivalent form. Eventually, it will assume "the character of direct and universal exchangeability" (161).

Commensurability through geometric quantification, striving toward universal exchangeability. Marx's understanding of the commodity starts to sound a lot like vector media. Neural networks don't just automate processes, they seek to create spaces in which all media are commensurable; their aim is precisely to produce *neural exchange value.*

We argued in the previous chapter that the distinction between compression and embedding is not deterministically technical, but ideological—as in the case of gzip, the "dumb" compression algorithm which in 2023 experienced a kind of epistemic metamorphosis into a text embedding. We can now complete that definition. Compression becomes embedding when data enters the equivalent form, that of neural exchange value.

Commodities, and their "metaphysical subtleties and theological niceties" have, of course, been at the center of many recent debates in historical materialism (e.g., Sprinker 2008), and we bring it up here not to revisit or rehash these debates, nor to steer our argument toward the important Marxist critiques of artificial intelligence, statistical reasoning, or mental labor in general (see esp. Joque 2022; or even Sohn-Rethel 1970). We wish, rather, to point out the fundamentally geometric nature of neural exchange value, which helps clarify why embedding has become the dominant paradigm for representing and manipulating knowledge.[2]

Neural exchange value offers not only a way to structure information, but a means to circulate it, exchange it, and repurpose it across domains. Media objects, of course, have been commodities long before neural networks—as, for instance, in Theodor Adorno and Max Horkheimer's (2016) account of the homogenizing power of the culture industries—but only in the sense of *market* exchange: Now they are becoming commodities in terms of *neural* exchange.

The double life of the commodity gives it an internal contradiction: It serves as "matter for two things at the same time," a "contradiction between the different forms of the concept" (Macherey 2016, 151). Commodities are indeed *defined* by this dual nature:

They always exist both in a "plain, homely, natural form" and as an equivalent form (Marx 1967, 138). *Computational commodities* are, too, characterized by their double nature: first, as texts, images, sounds, and *also* as neural embeddings.

Neural exchange value, then, is a new exchange form. In this light, each new neural network is not itself merely a product or a commodity in the traditional sense (to be sold, rented out, profited from), but rather a constructed space of exchange in its own right. Neural networks—and especially, as we shall see, multimodal foundation models—are less like clothes or automobiles and more like stock markets or currencies.

Naturally, the political and economic stakes of dominating the space of exchange far exceed those associated with control over individual commodities. It is precisely this heightened significance that renders exchange forms so alluring to Silicon Valley. This strategic turn was foreshadowed by the dynamics of platformization traced by Nick Srnicek (2017), though they remained tethered to traditional markets. Platforms mediate and monetize existing exchanges, such as matching drivers to passengers, rather than constructing autonomous logics of exchange. A cruder yet revealing parallel can be found in the exuberant proliferation of cryptocurrencies in the early 2020s: Here, the attempt to establish novel spaces of exchange manifested in the crude, and often unsuccessful, creation of new currencies—a gesture markedly less sophisticated than the exchange spaces constituted by vector media.

Distributed Representations

In the previous chapter, we saw that neural networks tend to produce embeddings that are seen as useful, even when trained on effectively epiphenomenal tasks: A network trained to distinguish between images of dogs and cats, for instance, might prove a reasonably successful search engine.[3] We might well ask for whom these embeddings are considered useful or meaningful, yet there

is no doubt that, technically speaking, deep learning embeddings have proven far more powerful than anything that preceded them.

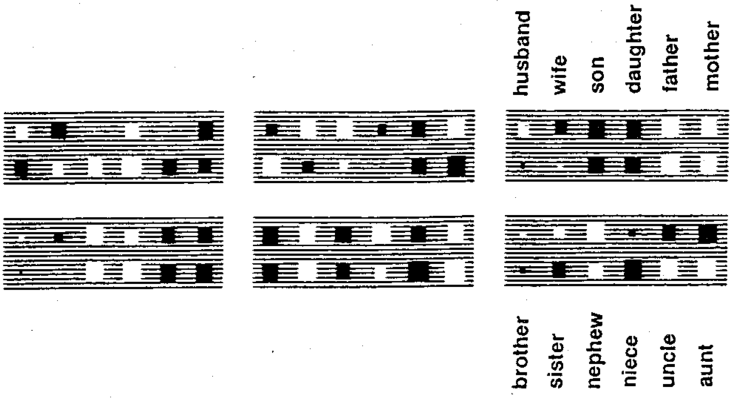
We have traced the intellectual history of the notion of embedding, from Graßmann's multilinear algebra to the operational embeddings of contemporary natural language processing. We also established that neural networks have become the paradigmatic machines for producing such embeddings. But this raises a further question: Why have neural networks, among all computational architectures, come to dominate this epistemic field? Why is it these systems, rather than any other, that have become our primary tools for producing relational knowledge?

To complete this story, we must turn to the work of Geoff Hinton, often described as the "godfather" of AI. Hinton, a Turing and Nobel prize winner, was a key player in developing a *probabilistic* mechanism for creating internal representations in computer vision. Hinton is well-known for, among other things, popularizing backpropagation in AI in 1986 (Rumelhart et al. 1986), the fundamental technique on which almost all neural networks since have been trained, and making it practical to train deeper (multilayer) neural networks for the first time.[4] Once again, we might say that the innovation of backpropagation is conceptual and ideological rather than strictly technical: It seems to have been mathematically invented independently at least twice before Rumelhart and Hinton's paper, but nobody had yet outlined its value for creating embeddings. As Hinton himself explained, "I have never claimed that I invented backpropagation. David Rumelhart invented it independently long after people in other fields had invented it.[. . .] What I have claimed is that I was the person to clearly demonstrate that *backpropagation could learn interesting internal representations and that this is what made it popular*" (Hinton 2020, our emphasis). What constitutes an internal representation, and what renders it "interesting"? Here, Hinton emerges as the pivotal figure who completes the trajectory traced in the preceding chapter: from compression to embedding. Hinton demonstrated that when data is forced through a neural bottleneck,[5] the resulting internal

representations capture patterns that abstract beyond mere input-output mappings. Unlike a file compressed in the traditional sense, these internal representations—which, as we saw in the previous chapter, we now call *embeddings*—do not simply memorize individual examples; rather, they encode relational structures.

Hinton often calls these embeddings "distributed representations." In the 1980s, the conventional way of representing information in a network was to assign one element of memory to each possible input. If text were represented via these "local representations" (now more often called "one-hot encodings"), for instance, each possible word in a fixed dictionary would be represented by a single element in the embedding vector, with all other values being zero. Debates over local versus distributed representations in neural networks were tightly bound up with the existence of "grandmother cells": whether the human brain might contain a single neuron that would represent one's grandmother, or whether the concept of grandmother would be distributed over neurons representing family ties, gender, age, certain facial features, clothes, spaces, smells, et cetera.

In this local system, each memory location was isolated, with no intrinsic relationship among words. Hinton, instead, was a major proponent for a *distributed* representation, one in which information about a stimulus is distributed throughout the vector in memory. The advantage of such a distributed representation is that it provides a continuous space, allowing exploration between existing datapoints—something impossible in the local (grandmother-cell) paradigm. A space, in other words, in which interstitial concepts can be extracted through interpolation: "[Comparing] distributed and local representations from the standpoint of creating new concepts. Local representations appear to require a homunculus with a soldering iron. [. . .] Distributed representations are more flexible, allowing concepts to be formed gradually by weight modifications that progressively differentiate one old concept into several new ones" (Hinton 1984, 3). Hinton is clear that many of his innovations (e.g., backpropagation) are not



[Figure 4.2]. Hinton's distributed representations. In this example, twelve neurons learn distributed representations of family relationships. Reproduced from Hinton (1986).

directly inspired by biological mechanisms, and in many cases go against them. Distributed representations, for instance, will be less sparse: a greater subset of neurons will have nonzero activations, on average, at any point in time. This goes against the logic of biological neural networks, which, in order to save energy, will try to minimize the number of active neurons at any one time—at least this is the intuition behind Olshausen and Field's experiments in sparse coding of natural images, for instance, which we saw in the last chapter. Hinton appears in this light less of a neuroscientist and more of an engineer, interested in the machinery of artificial, more than natural, intelligence. And yet, Hinton's revolutionary proposal is not strictly technical, but epistemic. It revolves explicitly around what we had identified, in the last chapter, as the defining ambition of vector media: the claim of being able to generate *new* knowledge.

Where did Hinton's interest in "interesting internal representations" emerge from? In his 1977 PhD work at the University of Edinburgh, Hinton was working on dealing with visual ambiguity in image recognition.[6] He was interested specifically in the problem of

"organizing the interactions between different hypotheses during the process of constructing the internal representation of a scene" (Hinton 1977, 9). Computer vision systems, Hinton argues, need to evaluate several potentially contradictory hypotheses in parallel, which might then be compared numerically in order to find the best one.

Just as striking as Marx's computer graphics analogy for exchange value is Geoff Hinton's (1977, 37) own analogy for the creation of a space of competing hypotheses in computer vision—the epigraph for this chapter worth reprinting in full: "In *Capital*, Marx puts forward the idea that there must be some common underlying essence shared by all goods in order to explain how they can be given prices according to which they are exchanged. The same philosophical point seems to apply to hypotheses [about digital images]. There must be some property which they share in order to explain how they can be given scores according to which they are traded. The obvious candidate is probability." If the internal representations, the vector embeddings, of contemporary artificial intelligence algorithms seem well described by our concept of neural exchange value, it is perhaps because the founder of modern artificial intelligence based the idea quite explicitly on Marx's notion of exchange value.[7]

We have seen in the last chapter that the notion of embedding has been—technically and epistemically—increasingly bound up with that of compression. But one of Hinton's major influences has been to shape the way in which that compression takes place. Compression is normally understood as a process of eliminating redundancy. From the point of view of Shannonian information theory (which, as the reader will recall, explicitly *excludes* questions of meaning), a perfectly compressed signal is one without redundancy. In Horace Barlow's account, the best neural representation would result from the smallest possible information bottleneck; this is the so-called efficient coding hypothesis. Hinton's view of compression is rather different; it takes the form of a *redundant bottleneck*.[8] This idea is most clearly exemplified in his invention

of "dropout," a technique that has become a standard method in deep learning. Dropout involves randomly "dropping out" (i.e., turning off) a subset of neurons during neural network training. By forcing the network to learn to make predictions without relying on any single neuron or specific configuration of neurons, dropout reduces overfitting and enhances generalization. It is effectively a form of what computer scientists call regularization, aiming to produce internal representations, which are more "robust." Dropout thus deliberately introduces a kind of redundancy into the compressed representation, effectively going *against* Barlow's original intuition. This approach reflects an important shift in the tacit theoretical approach of machine learning: Epistemic compression is perhaps no longer about minimizing redundancy in the Shannonian sense, but about strategically *managing* it.

If text has always been a model of symbolic reasoning, and images a model of perception, then these two things meet through compression, which, in turn, claims to model all knowledge. What is more, the fusion of reasoning and perception in compression attempts to open up a universal potential for generating new knowledge not within, but between, media (or, more precisely, signal domains)—knowledge that is neither text, nor image, nor sound. This is why Hinton (1986) can claim to be working not on images or texts but *concepts*—direct representations of knowledge. Distributed representations, he suggests, lead to "automatic generalization."

We will come back to the media-theoretical implications of this shift soon—but first it is important to emphasize how fundamental it really is. If early versions of embedded knowledge like Gerard Salton's document space were construed as essentially structuralist operationalizations, mappings of abstract document properties to vector space axes, Hinton takes us—quite explicitly, as it turns out—beyond this *structuralist*[9] paradigm and toward a paradigm that we might call *post-structuralist.* Obviously, the aim of what is now called post-structuralist philosophy could not be further from the aims of artificial intelligence research. We could easily

speak here of a technical rendering of *différance*, but, somewhat ironically, this would itself be another forced analogy enabled by embeddings. Instead, the similarity we would like to point out here is epistemic. Meaning and knowledge, for Hinton, are to be found in the gaps between the representations, not in the representations themselves. These gaps, as the case of word embeddings makes so abundantly clear, come into existence only because continuity is added to the representations—a potential for being more than they are. As Hinton (1990, 58–59) writes, building explicitly on Barlow:

> These two extremes are the natural implementations of two different theories of semantics. In the structuralist approach, concepts are defined by their relationships to other concepts rather than by some internal essence. The natural expression of this approach in a neural net is to make each concept be a single unit with no internal structure and to use the connections between units to encode the relationships between concepts. In the componential approach, each concept is simply a set of features, and so a neural net can be made to implement a set of concepts by assigning a unit to each feature and setting the strengths of the connections between units so that each concept corresponds to a stable pattern of activity distributed over the whole network.

Crucially, Hinton's ambition *requires* a fundamental commensurability between embedded signals—a medium of exchange, so to speak. If what we are looking for is to be found in the space between representations, then the original ontological properties of those representations—what makes an image an image, a text a text, the thing in itself—cease to matter.

Yet the fundamental commensurability enforced by embeddings is only one side of the ideological coin. The other is the desire to explicitly transcend the epistemic boundaries of traditional machine learning and argue for its ability to make *new*

knowledge—machines that, in Hinton's (1986) view, are intended to produce knowledge by interpolating between the "representations of concepts." The claim is that generalization becomes a form of cognition. This is where the (materialist) critique of artificial intelligence must meet the (materialist) critique of automation. As Matteo Pasquinelli (2023, 244) points out, artificial intelligence emerged historically "from the automation of the psychometrics of labor and social behaviours rather than the quest to solve the 'enigma' of intelligence." Neural networks extend this logic to its extreme point; they do not merely measure, but encode abstract equivalence, an equivalence that *used* to be tied more or less concretely to labor, as a form of operational value.

Beyond Quantification

In the previous chapter, we noted Jonathan Sterne's call for a cultural history of compression. Compression, for Sterne, is not a mere technical protocol for an impoverished derivative of some unspoiled original, but rather an operation with historical force—a cultural technique. What would it mean, Sterne asks, if compression replaced representation as the dominant axis of inquiry in media and aesthetic theory? Alexander Galloway and Jason LaRivière (2017, 126) have since taken up Sterne's provocation, aiming to "reexamine representation within the Western philosophical tradition" through the lens of compression. Their framework differentiates two main types of philosophical compression: *abstract* compression, which views compression as an "undesirable by-product of the metaphysical contract," where the "real" is uncompressed and everyday phenomena are "selective deletions of a superlative nature"; and *generic* compression, involving "data deletion at the level of real material life," which results in a "positive tactic of material indifference" (127).

Through this expanded lens, Galloway and LaRivière reread key figures in media theory and philosophy as theorists of compression *avant la lettre.* Friedrich Kittler, for instance, identifies literature as

the reduction of speech into a finite symbolic system—an "alphabetic monopoly" channeling all data through the bottleneck of the signifier (Galloway and LaRivière 2017, 126). Alain Badiou positions fidelity as compression's inverse: a sustained commitment to an event that resists reduction. Bernard Stiegler, building on Jacques Derrida and Sylvain Auroux, theorizes grammatization as the discretization of continuous experience into manipulable media units. Yet for Galloway and LaRivière, it is Sterne who pushes furthest: not toward recovering what has been lost, but toward accepting compression itself as the dominant condition of mediation.

Alongside compression, we might look for a history in the *longue durée* of embedding. Adrian Mackenzie (2015, 432), writing several years earlier, is interested in practices of "generalization through vectorisation." In *Machine Learners* (2017), he explores vectorization as a practice that restructures how knowledge is modeled and enacted ("What is at stake in vectorizing data?"; 73)—which we have been calling epistemic compression. In this new spatial logic, identities and differences no longer rest on resemblance or shared origin, but on calculated proximity. "It produces a common space that juxtaposes and mixes complex localized realities" (73)—images and text, for instance. Once again, for Mackenzie, the ideological undercurrent is implicit in the process of vectorization even before any notion of machine learning. Historically (and with Michel Foucault), vectorization, as an epistemic practice, is what follows after the table, the former "dissolving" the latter: "Tables, we might say, become merely data, inscriptions dependent on hidden structures and their genesis." (57). Referencing Mike Lynch, he names this effect "epistopic" (59): the imposition of generalized epistemic forms onto domains that resist generalization. It is a fault line rather than a synthesis—unstable, yet productive.

Faced with the long historical arcs of computation and quantification emphasized by Galloway, LaRivière, Mackenzie, Kittler, and others, we might suspect that the very notion of neural exchange value (or, more generally, vector media) is superfluous. Are contemporary neural networks not simply another case of

measurement, another instance of what computers (among other technologies) have always done through discretization, classification, or more generally digitalization?

More specifically, part of what makes vector media appear continuous with earlier computational regimes is their inheritance of the erasure of the boundary between code and data first introduced by Alan Turing and John von Neumann.[10] Once instructions could be stored and manipulated like any other data, the computer became a medium agnostic to this division between instruction and raw material, between form and content.

Yet, to read vector media only against these continuities is, unfortunately, rather too easy a solution. The legitimate critiques that have been leveraged against these cultural techniques and their many media-technological operationalizations are not in themselves sufficient to understand the new dynamics of deep neural networks. In fact, as briefly discussed in the first chapter, many popular general critiques of artificial intelligence fall back to a discussion of the homogenizing force of these techniques. Embedding undoubtedly shares properties with all of them, not least because it relies on the computer as its carrier. One could legitimately write a Kittlerian archaeology of inscription in which vector media find their place as the latest phase of an old regime. Such a history might well surface fascinating continuities, but it would also miss some key material specificities of our present condition. Vector media have other long historical continuities—such as in the history of philosophies of visual perception, as we have attempted to sketch in the second chapter. Yet the dynamics of vector media are *not* the determining phenomena of the abacus, the Pascaline, the Jacquard loom, ELIZA, or the GOFAI script in LISP.

Vector media simultaneously amplify and transcend the epistemic and ideological limitations of computation, then, in ways that demand separate theoretical consideration. Neural embeddings do not merely measure culture; they construct an entirely new regime of value, one that defines meaning through patterns of similarity

and difference rendered legible to computation. To clarify how vector media both intersect with and depart from these foundational techniques, we now turn to consider each, moving down the stack in turn.

Classification: Critiques of artificial intelligence often rest on the assumption that AI research is inherently bound to an ideology of classification (Bowker and Star 2000). The well-known flaws of datasets like ImageNet are frequently cited as evidence of epistemic failure, suggesting that the structure of a dataset determines the epistemology of any model it produces. Yet such critiques miss a vital distinction: Challenges like ImageNet were never merely attempts to refine systems of classification, but efforts to transcend them entirely. To imagine that machine learning models are forever shackled to predefined labels is to overlook the radical possibilities that embeddings introduce. Unlike the rigid taxonomies of GOFAI, vector media dispense with fixed axes altogether. They operate as one-way processes in which embedded objects are not recoverable in their original (natural) form, and the relationships between them thus simply *cannot* depend on fixed taxonomies or categories. And, in fact, as Thao Phan and Scott Wark (2021) and others have compellingly argued, it is precisely this collapse of rigid hierarchies that grants corporate actors their unsettling precision, allowing them to exploit the ambiguity inherent in vector media to construct a perverse intersectionality that exceeds the limits of conventional classification.

Digitalization: This is perhaps the most pervasive misunderstanding: How does embedding, especially in its power to render disparate media forms apparently commensurable (more on that later), not simply mirror the encoding of these media forms as digital data? Is it not all "ones and zeros" beneath the surface? Aside from the obvious counterexamples of optical neural networks, this is a fundamental misunderstanding of the nature of embedding, which, compared to something like traditional digital file formats, is an incomparably narrower, more radical, and more violent form of encoding. A digital image, for instance, is still an image in the sense

that the relationship of sensory data and pixel data is comprehensible, linear, and deterministic (as, in the extreme case, is the relationship between letters, phonemes, and words). This is not at all the case for embedded images, which represent sensory data in an exclusively relational form that is simply incomprehensible outside the scope of all other images that a model has also ingested. In vector media, neural exchange form suspends the possibility of reconstructing the original sensory data entirely. There is simply no equivalent in the realm of digitization to this phenomenon. (Almost) all vector media are indeed digital, but the goal of every form of traditional digital encoding is reconstruction, communication, and computation. Vector media invert this logic: Embedding departs from reconstruction as a goal, producing representations that exist only in relation to other embedded forms. In many cases (e.g., autoencoders), reconstruction is pursued only as a route to learn an embedding space. In practical terms, this creates a profound separation. The embedded image acquires an autonomous existence, haunting the sensory data from which it arose, data that must now be stored elsewhere, connected to its embedding only by a fragile thread—a database link or an array index.[11]

Finally, *discretization,* the more general case of digitalization. Here, we have arrived at the bottom of the stack (in a technical, not Brattonian, sense), where, supposedly, the technical becomes ontological. Isn't embedding, we might ask, simply another way of sampling, of forcing a fundamentally continuous reality into a discrete system? Consequently, aren't the ideological implications of embedding not simply the ideological implications of discretization—that is, a loss of continuity, and thus a loss of ambiguity—and thus, eventually, a kind of violent quantification à la Adorno and Horkheimer (2016)? We could call this the *cuneiform hypothesis*: that ontologically, AI can be understood as no more than a logical and inevitable extension of discretized glyphs. We would like to suggest that assuming this position leads to an extreme reductionism that actually forecloses any meaningful critique of artificial intelligence.

The conflation of embedding with discretization mistakes a structural affordance for a political position. While embedding certainly presupposes discretized infrastructure—bits, floats, numerical arrays—the politics of vector media cannot be reduced to questions of resolution or bit depth. Compression, as we understand it here, may include technical operations like quantization or reduced precision, but these are not its epistemic core. The salient feature of embedding is not that it coarsens data as such ("pixellisation," as in Steyerl's poor image), but that it models continuity synthetically, through high-dimensional structures designed to simulate gradience, ambiguity, and nuance. This becomes clear if we return to early work in neural networks. Both Hinton and Kunihiko Fukushima phrased their seminal neural architectures not as computer algorithms, but as analog, continuous systems—their computational results presented as *simulations* of the neural network,[12] not the network itself. If the politics of AI were truly located in its discretization—its float precision, its quantization level—then it would follow that neural networks might become less ideological, or less violent, simply by upgrading to, say, 64-bit double precision floats. This is clearly absurd.[13] The ideological work of embedding happens at another level entirely: not in the digitization of signals, but in the construction of relational spaces that present themselves as natural ontologies.

We can see examples of such an extreme reductionism in (some of) the writings of Friedrich Kittler (2012), who famously argued that "there is no software,"[14] or, more recently, in Angus Fletcher's critique of artificial intelligence from the perspective of literary studies. There, Fletcher (2021, 1) argues thus: "(1) Computer artificial intelligences (including machine-learning algorithms) run on the CPU's Arithmetic Logic Unit, which performs all of its computations using symbolic logic; (2) symbolic logic is incapable of causal reasoning; and (3) causal reasoning is required for processing the narrative components of literature, including plot, character, style, and voice." Apart from the fundamentally incorrect statement about CPUs (such algorithms can run not only on GPUs, but also,

for instance, in entirely optical systems), Fletcher is clearly arguing for a preservation of epistemological affordances from the bottom up: It is the discrete ("symbolic logic") nature of computation that forecloses higher-order ("causal") reasoning.

Ironically, such an extreme reductionism very much practices what it critiques: A technical system, in this view, is either bad (symbolic) or good (literary). Yet the key ideological mission of vector media is to exactly *overcome* the limitations of discretization, which had plagued both GOFAI and "local" vector representations: to model the ambiguity missing from its representations by adding multidimensional *continuity*. In fact, vector media does this even in cases where such continuity is clearly contradictory, such as in Hinton's (1986) work on family relationships (e.g., niece, uncle, father, brother), or more subtly in word embeddings: Neural exchange is created precisely through the artificial interstitial continuity of inputs, which can clearly only exist naturally in discrete form.

Embedding, as we argued throughout, is itself a cultural technique—*the* cultural technique of vector media. While explicit embedding layers exemplify its logic, the cultural scope of embedding far exceeds any single layer or architectural block. The penultimate layer of the obsolete ImageNet-trained VGG-19 model becomes an embedding only when we instrumentalize it as such; more complex architectures, such as Stable Diffusion, do away with dedicated embedding spaces altogether.[15] What remains is the general form: the one-way compression of organic media into quasi-continuous, distributed, exchangeable representations. This is the operative technique that renders almost any machine learning system susceptible to embedding. And it is this pervasive conversion—from the particular to the relational, from the symbolic to the geometric—that marks the broader terrain of vector media. Just like every media object can be sublimated into vector media, (almost) every machine learning object can itself become an embedding space, because the underlying cultural technique pervades machine learning in a much more general sense. Almost all machine learning systems, then, have become vector media.

Vectors and Language

Exchange value, Marx insists, is fundamentally relational. It exists only relationally—it does not inhere in individual commodities—precisely because it emerges from the dynamic comparisons between them. The same applies to vector embeddings in vector media: Neural exchange value is generated not by any absolute measurements but by geometric relations, hence the recurring conceptual importance of similarity in neural networks.

Language, too, is fundamentally relational. This is the basic assumption of distributional semantics, as in J. R. Firth's proposal (following Wittgenstein) that a word is defined by the company it keeps. Distributional semantics is of course a founding assumption of many NLP models, such as word2vec, which, as we saw in the last chapter, were themselves heavily influential in developing the ideology of vector media. Embedding models take this literally: Linguistic (or rather semantic) meaning is a function of geometric proximity within a vector space. But language is relational in a wider sense, in that it is fundamentally a social activity.

It is precisely at the discordant intersection of these two forms of relationality—social and semantic—that we can start to see a notion of exchange value in forms of linguistic exchange. We can see a precedent for this kind of understanding in Marx himself who, in the *Grundrisse*, compares exchange value not to language in itself, but rather to another key conceptual proving ground for vector media, translation between languages: "To compare money with language is not less erroneous. Language does not transform ideas, so that the peculiarity of ideas is dissolved and their social character runs alongside them as a separate entity, like prices alongside commodities. Ideas do not exist separately from language. Ideas which have first to be translated out of their mother tongue into a foreign language in order to circulate, in order to become exchangeable, offer a somewhat better analogy; but *the analogy then lies not in language, but in the foreignness of language*"

(Marx 2005, 67, our emphasis). This is to say, ideas can only exist within language, but that does not mean that they are dissolved into language. Translation thus is an ambivalent process: It enables the circulation of ideas, amplifying their sociopolitical efficacy, but also means that they become "more" language and "less" idea.

In "The Task of the Translator,"[16] Walter Benjamin (2002) famously notes that it is translation where the historicity of language becomes most apparent, it is where language becomes somewhat separated from the realm of concepts which Benjamin calls "pure" language: "In this pure language—which no longer means or expresses anything but is that which is meant in all languages—all information, all sense, and all intention finally encounters a stratum in which they are destined to be extinguished" (261). We rephrase Benjamin's point in the following terms. (Neural) exchange value serves to obscure both its natural form (obscuring also the historicity of that natural form, such as language) and the labor that was expended to extract it. This process is always incomplete: Neural exchange value is always frustrated by the epistemic incommensurability of the natural form.

Many others have drawn analogies between language and exchange value. Karen Bennett (2022) reminds us that Saussure (in the *Cours de Linguistique Générale*) speaks of translation in terms of the political economy, with figures as diverse as Eugene Nida and Cicero giving the same metaphor between currency-exchange and translation. Ferruccio Rossi-Landi (1992, 83)[17] wrote that a "linguistic community works like a sort of vast marketplace in which words, expressions and messages circulate like commodities." Franco "Bifo" Berardi (2012, 6–7) speaks of the "automation of language" leading to a "reduction of language to exchange," and therefore that the "passage from the industrial abstraction of work to the digital abstraction of work implies an immaterialization of the labor process" (134–35). David Graeber (2001) connected the relational aspect of language to questions of value, placing linguistic value in interrelation with moral and economic forms of value.

Such analogies are all drawn with what computer scientists would call "natural" language, yet contemporary vector media stretch the metaphor of translation yet further: Free text is translated into computer code (as in the recent phenomenon of "vibe coding") or musical scores (which are most often notated symbolically in machine-readable markup, e.g., MusicXML). The limit case here is the use of large language models for the prediction of protein structure (Offert et al. 2024). How exactly can we understand the "neural exchange value" of this?

Media Collapse

Contemporary AI, then, reveals something far more fluid than discretization, digitization, or classification: a transformation in the very relationship between form and content, syntax and semantics. Embedding space is, fundamentally, a *medium* in which to compare anything with anything else, a medium of the equivalent form—not unaffected by, but rather intentionally indifferent to, the source form of its raw material. Unlike image files and sound files, which refer more or less explicitly to photographs or cassette tapes, embedding space is a medium that exists only *inside* the black box. It is *operational*, in Harun Farocki's sense,[18] in that it is conceived not for direct human consumption, but rather to facilitate subsequent machinic manipulation and processing (e.g., neural exchange): Embeddings fuel search engines, classifiers, generation networks, and so on.

The promise of embedding is that everything can be made comparable. Text, image, sound, even protein structures—all can be projected into the same abstract space. If the previous chapter traced how meaning emerges not from the intrinsic content of media objects but from their position in a space of relations, then vector media radicalize this premise by extending it across modalities. Embedding no longer simply describes linguistic proximity or visual similarity; it becomes the mechanism through which fundamentally different types of signals are rendered interoperable. Vector media establish a space of general commensurability in which meaning

is produced through exchange, rather than representation per se; this is the promise of highly multimodal artificial intelligence, of foundation models,[19] and ultimately of artificial general intelligence. What matters is not what something is, but where it is, and how it relates to everything else.

If embedding spaces are spaces for the neural commodification of media, then this phenomenon can be understood as a kind of *financialization* of media. Just as financial markets have increasingly subordinated the "natural" economies of goods and services to abstracted systems of exchange, model collapse suggests that the embedding—the vectorized abstraction of meaning—comes to dominate over the "natural" form of language, images, or other data. The natural form becomes subservient to the exchange form, as embeddings operate as a universal currency of meaning.

Universal, because exchange forms always attempt to become the dominant equivalent form. This logic is not incidental—it is programmatic. The ambition started with so-called foundation models within individual modalities of data: ImageNet famously aimed to *map out the entire world of objects*,[20] while Stable Diffusion, as we saw in the last chapter, claims to compress the *visual information of humanity* into a shared vector space. But the shift to multimodality is not simply an expansion of the kinds of input files with which the network is expected to deal, a mere convenience: it is the ambition to model *thought*, as the central epistemic infrastructure of machine reasoning. Geoff Hinton himself speaks of *thought vectors*: "You can capture a thought by a vector" (Devlin 2015). In a 2015 keynote at the central AI research conference CVPR, Yann LeCun (2016, slide 106), then chief scientist of Meta (formerly Facebook) AI, took the notion even further: "Reasoning consists in manipulating thought vectors. [. . .] At FAIR [Facebook AI Research] we want to 'embed the world' in thought vectors. We call this World2vec." There is, then, a structural ambition to collapse every possible input—whether word, image, protein, or event—into a single medium of *cognitive* equivalence. If the media theory of the 1990s had to discover the computer as a medium (Bolz et al. 1994),

today's media theory has to grapple with the fact that all media are dissolved in the modeling of cultural artifacts with embeddings: a total *media collapse.*

This totalizing ambition is already visible in the technical literature. Ledell Wu and colleagues (2018) proclaimed the imperative to "embed all the things," laying the groundwork for many subsequent general-purpose architectures capable of embedding heterogeneous modalities of data into a shared embedding space. More radically still, Hao and colleagues (2024) attempt to perform reasoning itself not within language tokens, as a conventional chatbot, but within the continuous vector spaces of the machine—reframing cognition as vector manipulation. Thought vectors here serve not only to *represent* or *generate* thoughts, but to *replace* them—to do away with linguistic cognition entirely. As we saw in the last chapter, this irreducibly vectorized cognition has important consequences for both technical interpretability and humanistic critique.

The paradoxes of universal translation through a geometry of equivalence surface in the form of technical constraints. Many architectures rely on textual captions as a common anchor—but alternatives exist. ImageBind (Girdhar et al. 2023), for instance, uses images to align text, audio, depth, thermal data, and movement (IMU) into a shared embedding space. Yet this collapse is never total. As Victor Liang and colleagues (2022) show, when embeddings cluster by modality rather than converging, the performance and generalization of the model suffer drastically. The very contradiction that makes vector media operational—their indifference to media—also exposes their limit: They can only know the world by erasing the form it came in.

We would like to warn, however, against seeing these contradictions as heralding an inevitable collapse of neural exchange value, vector media, and the entire political and technical ecosystem of artificial intelligence. The critical vocabulary surrounding AI has begun to echo a strangely familiar register, one that Marx and Engels once reserved for the self-destruction of capitalism. AI, like capitalism, is said to produce its own gravediggers:

Terms like mode collapse and model collapse are taken up with a certain apocalyptic relish, often by those only loosely tethered to their technical meanings. But this historical irony masks a deeper problem. Collapse, in this discourse, becomes not diagnosis, but fantasy, in which technical failure stands in for political transformation. These failures do not unravel the system—they deepen its hold. We must resist the lure of collapse as catharsis, the romanticization of glitch as political rupture. Unfortunately, as we will see, the failures of AI do not mark its unraveling—instead, they signal its deeper entanglement in the systems of production, moderation, and control.

First, a technical clarification on the current eschatological discourse is in order. *Mode collapse* describes a failure mode of generative adversarial networks, wherein the generative model, in seeking to minimize its loss function, learns to reproduce only a fragment of the training data's distribution. It is a failure that, especially in image models, reveals itself dramatically: A model trained on handwritten digits might lose itself in the contours of the zero and the one, forgetting entirely the rest of the numerical spectrum. The collapse is stark, catastrophic, and easily visualized—a narrowing of diversity to a single, repetitive kernel.

Model collapse, on the other hand, occurs when AI models[21] are trained on data that includes the output of previous models—something seen as likely for models trained on large quantities of images or texts from the internet (Shumailov et al. 2024). "Model collapse" was formally characterized in 2024, but something analogous had been noticed over a decade earlier, when researchers from Google noticed that automatically translated text made up a significant proportion of multilingual text on the web: "As new services that output structured results are made available to the public and the results disseminated on the web, we face a daunting new challenge: Machine generated structured results contaminate the pool of naturally generated human data" (Venugopal 2011, 1363). However, model collapse does *not* signify the dramatic implosion of AI systems, as might be imagined by analogizing it

with mode collapse in generative adversarial networks. Mode collapse represents a catastrophic failure mode, where the system effectively ceases to generate meaningful diversity in its output, restricting itself to a narrow subset of its input distribution. In contrast, model collapse entails something far more insidious: a progressive smoothing of the outputs, marked by a reduction in complexity and unpredictability. It is less a dramatic collapse (Zhang and Pavlick 2025) than a kind of entropic drift toward mediocrity—a narrowing of the possibilities of expression as the system becomes more predictable, less "glitchy," and more uniform in its patterns.

In technical terms, this can be measured as a reduction in *perplexity*, a key metric in information theory (often used in language modeling) that captures how well a model predicts a sequence of data. As models consume their own outputs in training loops, the diversity of their generated outputs diminishes, creating a feedback loop of homogenization. Language becomes dominated by statistically probable but less imaginative constructions. The unexpected artifacts that scholars and artists working with AI have come to appreciate (and which corporations tend to fear) are gradually smoothed out.

This predictability is not merely an aesthetic phenomenon, but one with broader political and cultural ramifications. The increasing blandness and banality of AI-generated content (sometimes called *AI slop*) aligns with the pressures of moderation and censorship, particularly in image- and text-generation systems. Content moderation protocols, often built into these systems to prevent the generation of offensive or harmful material, further contribute to a flattening of possibilities (Offert 2023). As political, social, and cultural sensibilities are encoded into models, they create a form of censorship that prioritizes conformity and predictability over dissent. Far from a utopian self-destruction of AI, then, model collapse marks a shift toward a homogenized, financialized regime of production of information commodities.

Marshall McLuhan famously argued that media are always defined in relation to other media. A new medium is always an extension, amplification, or transformation of an existing one. Yet in the era of embeddings, this relationship is fundamentally altered. Vector media are a kind of violent meta-medium, capable of representing all media through the universal language of vector embeddings, of neural exchange value. Vector media, then, dissolve all others. We call this phenomenon *media collapse*: It represents a far more profound and structural transformation than the relatively bounded technical phenomena of mode collapse or model collapse. To understand this collapse in practice, we can look to OpenAI's CLIP (Contrastive Language–Image Pretraining) model—a paradigmatic example of how embeddings reshape the relationship between media.

Media Collapse in Practice

CLIP is explicitly designed to bridge modalities, creating a shared embedding space for both text and images. Unlike models that incidentally generate embeddings while being trained on other tasks, CLIP's architecture is intentional in its goal: to render visual and textual data interoperable by encoding them into a common vector space. In doing so, it reduces the affordances of both media to a kind of lowest common denominator, flattening their specificity to achieve a computationally legible form of commensurability.

Naturally, CLIP fights an impossible battle. Certain words (e.g., *red*) are always going to translate more easily into visual concepts than others (e.g., *however*). Similarly, some visual phenomena evade textual description: Michael Baxandall (1979) notes that European languages tend to discriminate finely in certain areas, such as Euclidean form, while remaining coarse in others, such as surface texture—an observation that could be made about many current-generation architectures of visual models as well, as research on so-called shortcut learning has shown.

And yet, CLIP performs remarkably well. Its multimodal embedding space allows it to create a relational understanding of visual and textual data, enabling associations that go beyond fixed categories or simple descriptive tasks. Consider the following perspective on the Museum of Modern Art, New York's collection through the lens of CLIP. As part of a previous series of experiments (Impett and Offert 2024), we created embeddings for each work for which images are available on the museum's website. While the resulting corpus of roughly eighty thousand images is certainly not *complete* (as works that are not photographically documented are missing, and as the "virtual" collection is only a subset of the institution's physical holdings), it nonetheless provides a representative sample. With the help of a tool, imgs.ai (Offert and Bell 2023), built to facilitate visual search in large cultural heritage collections, we can utilize CLIP to search for the images "closest" (in embedding space) to a specific textual prompt. And it is here that CLIP's capabilities become apparent. A CLIP search for *rhythm*, for instance, will retrieve images that embody a large spectrum of the meaning inherent in that word: 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 somehow rhythmic. While contextual ambiguity, famously, became the most prominent roadblock of traditional word-embedding techniques, it appears to be "solved" in the multimodal space of CLIP, where the whole semantic field associated with *rhythm* is on display.

This ability to manage abstraction is partly due to CLIP's design as a bridge between modalities, where it "learns" visual concepts through textual descriptions and textual ideas through their visual counterparts. It's important to emphasize how foundational cross-modality is to CLIP's training: It learns about images *only* through textual descriptions, and vice versa. In doing so, CLIP embodies a kind of ekphrastic imagination, totally incapable of visualizing the unsayable (or parsing the nonvisual). Nonetheless, it *is* remarkably sophisticated in approximating paravisual ideas, operating in

spaces that are neither fully visual nor textual, but embedded in cultural associations. In previous work, we have shown how CLIP deals with (historically, philosophically) entangled concepts in art history. What we find, for instance, is that some—but not all—of the discourse surrounding the contested distinction between "naked" and "nude" can be mapped to CLIP's vector space, while CLIP's idiosyncratic preferences (for photography over painting, for instance, mirroring the preoccupations of computer vision) significantly impact the legibility of established discursive axes. In a perhaps even more striking example, we find that CLIP's concepts of cities—for instance, Paris—are highly spatially uneven, and mirror, quite uncritically, the social stratification of these cities. Paris looks like Paris, to CLIP, only in the center—while Johannesburg looks most Johannesburg, to CLIP, in its most poverty-stricken parts.

The interplay between specificity and generalization is foundational to CLIP's architecture: While it depends on medium-specific affordances like the spatial hierarchies of images or the sequentiality of text, it ultimately dissolves those distinctions into a shared embedding space. The result is a system that encodes visual and textual cultural imaginaries while simultaneously flattening their complexity. This tension—between medium specificity and medium indifference—is at the heart of media collapse.

Media collapse becomes even more striking when we consider the shared architectural lineage of models across modalities. It is not just that images and words are made comparable; the very architectures that enable this comparison are themselves caught in a similar contradiction. Neural networks, as we have seen, are often grounded in biologically inspired simulations of individual modalities of perception. Convolutional networks, for example, were originally conceived as models of the mammalian visual cortex, while the bottleneck structure of encoder–decoder networks draws directly from the physical constraints of the optic nerve. And yet these architectures now operate promiscuously across media. Deep learning architectures for text, for instance, borrowed heavily

from those developed for vision in the mid-2010s: Pioneering work by Yann LeCun and others demonstrated that convolutional networks could be adapted to process text as a kind of raw signal. LeCun and colleagues explicitly described their architectures—the first very deep convolutional networks in text processing—as a translation of computer vision paradigms (Conneau et al. 2017).

This cross-pollination of architectures extends beyond images and texts. Music, too, becomes legible to computational systems through a reduction to representations borrowed from other modalities. Musical notes (represented in XML, or as MIDI files) are often processed through architectures that are essentially borrowed from natural language processing. But how is raw audio processed? One frequent solution has been to transform the audio into an *image* (by taking a spectrogram), and then use computer vision networks, once again modeled on retinal neurons (see, e.g., Donahue et al. 2018). Media collapse, here, names a double movement: a collapse at the level of data, where distinct modalities are dissolved into neural exchange value; and a collapse at the level of architecture, where models built to simulate specific forms of biological perception are abstracted into interchangeable templates, applied across media without regard for their organic phenomenological grounding.

Many more cases of media collapse could be listed here. To misquote Louise Amoore (2024), to the computer scientist, all space is vector space—stretching from the quark to the nation-state, from climate models to poetry. It is a vision not of universality, but of assimilation: a topological imperialism where difference becomes a matter of scale, not kind.

One of the most important critical openings we see is that the parallelism and fundamental *incompatibility* of vector spaces quietly contradict their claim to universality. The debate around foundation models has successfully obscured the fact that the centralization of model development is, just like the dissolution of culture into neural exchange value, not only ideologically but technically

inconsistent. At any moment, multiple foundation models are competing with one another, and the market crash introduced by the release of the Chinese DeepSeek model tells us that this situation is unlikely to be resolved. There is, in other words, no unified "latent space of culture," as Ted Underwood (2021) has suggested in one of the first critiques of vector space hegemony. Not only are individual models necessarily limited in their training data, but the economy of neural exchange value has no gold standard, no ultimate reserve, no promise of exchange. One cannot convert GPT embeddings into DeepSeek: Spaces of exchange are both universal and mutually exclusive.

Even on the technical level, we find many operational admissions of defeat. Large visual models (have to) develop their own "emulations" of media, which, for instance, allow the quite precise setting of photographic parameters (f-stop, etc.) in image generators. What is more, the same kind of "conservation" happens also on the architecture level, where smaller (and older) models are often integrated into the processing pipeline of a larger model. The specters of previous models, in that sense, haunt the AI systems of today: The VGG-19 model is still responsible for the computation of the "perceptual distance" in many contemporary large visual models.

It is precisely "where the model meets the machine" (Dick 2015, 630) that the ideology of neural exchange value begins to falter. This is mirrored in many (desperate) technical attempts to render the perceived universality of vector spaces into reality—and even more prominently in broad philosophical claims, such as the recent "Platonic Representation Hypothesis" (Huh et al. 2024), which claims that, as models scale, their internal representations converge on a shared statistical form. This might just as well be the case—but certainly not because of an underlying universal epistemic structure that is being modeled, a platonic idea, as the hypothesis would have it, but because of the tendency to utilize ever more similar architectures on ever more similar data. In fact, models trained on different data, for different purposes, with different architectures, do not converge, or do so only under carefully

managed conditions. The "Platonic Representation Hypothesis" and its many quasi-empirical operationalizations, then, do not prove the universality of vector space, but they expose, once more, the deep desire for universal exchange that structures the ideological undercurrent of vector media.

Generativity Does Not Exist

We began this book with the problem of generativity. The conversation around so-called generative AI, particularly in relation to art, has remained oddly static. It circles endlessly around questions of authorship, originality, and creativity, as if the twentieth century had not already exhausted them, in discussions of electronic art and elsewhere. Too often, it fails to grasp the specificity of the machines we now confront: in other words, without a materialist analysis of what is specific to *our* machines.

Our attempt, following Agre, has been to start from a consideration of what has changed on the ground. To take seriously the technical and epistemic mutations that structure contemporary machine learning—neural embedding and epistemic compression—and to understand them not only as statistical operations, but as cultural techniques. These meet in contemporary deep learning networks, the ideal type of vector media. In this view, embedding spaces are not only spaces of representation, as is commonly assumed, nor of epistemic generation, as many central AI scientists would claim: They are spaces of exchange.

In truth, even the term *generative AI* itself is misleading. It treats generation as the defining feature of machine learning systems. Historically, the opposite is true. Generative architectures emerged not to create outputs, but to produce embeddings. The capacity to generate was useful, but largely in service of another aim: to shape a space in which vectors become comparable, exchangeable.

This structure is a consequence of what we described in the third chapter as the ideology of epistemic compression: the idea that the

best way to learn useful representations is to eliminate (or rather manage) redundancy, an intuition Horace Barlow came to when considering the optic nerve.

The obvious way to do this is to force information to "flow" through a bottleneck in a neural network. Take the autoencoder. It learns by reconstructing its own input—not because the output has intrinsic value, but because the act of reconstruction compels the network to discover a more compact internal representation. That representation becomes the embedding.

Even something as basic as PCA works in this way. Its components allow for a low-rank approximation of the original input—as in the case of eigenfaces, where the goal is not to create a perfect likeness, but to reduce the image to a manageable set of axes. Eigenfaces were used in person recognition (using the embedding), not to generate realistic faces for Hollywood. Once again, generation is a by-product of epistemic compression, ultimately in the service of neural exchange.

The same is true even for language models, the central case of so-called generative AI. In the last chapter we saw word2vec. One variant is the so-called skip-gram model, in which the task is to predict a missing word from its surrounding context: two words on the left, one missing word, two on the right. The network learns to generate plausible fillers for that central slot. But nobody uses those outputs—it would be hard to find many useful applications for software whose only task is to autocomplete for the middle word of a five-word phrase. What matters, instead, are the internal embeddings that emerge in the process—a vector space that encodes statistical relationships between words. Even large language models, which generate text by definition, are built on the same principles. During training, they produce billions of words that will never be read. These outputs exist only to sharpen the internal spaces in which prediction happens. Generation follows embedding, not the other way around.

What we have tended to call AI image-generation systems, too, are characterized by embedding and exchange. The most popular open-source image generation network, Stable Diffusion, does not generate images in a vacuum; it is guided by CLIP, the algorithm described above, which is a kind of ideal case of neural exchange value. CLIP does not produce images or texts, nor does it ask what an image or a text is; it only asks how close *this* image is to *this* text. Its function is not representation, but alignment.

We have come to the most radical consequence of neural exchange value. Contemporary generative artificial intelligence is seldom truly generative. They are, for the most part, systems of exchange rather than creation. They are spaces of vectorized commensurability. Any image or sentence produced by them is not created in a vacuum; it is extracted from a set of calculated proximities. The generative act is a traversal of embedding space, a movement from one vector to another. Generation is merely the moment when exchange becomes visible.

Prompting, then, is not ekphrasis; it is not an act of description. It is a transaction. A prompt selects a position in a space already defined by countless other vectors. It does not create, but retrieves—or more precisely, it activates the machinery of substitution that underlies the embedding space. What matters is not the specificity of the prompt, but its ability to produce a result that appears coherent, polished, complete—hence the baroque new dialects of image-prompts ("portrait of a teapot, 85mm f/1.4, ISO 200, soft lighting, Fujifilm Pro 400H, high detail:1.2, RAW"). The prompt is abstract labor made into tokens; the image is a fabricated surplus.

This is no less true in the realm of language models. A one-line request for "an email to apologize to a colleague" returns three polished paragraphs of corporate empathy. "Summarize this book" returns an object optimized for traction. These are not acts of generation but exchange, and more specifically arbitrage—exchanges in which minimal, compressed signals are "traded up" for the simulation of greater embedded labor.

Questions around whether AI-generated images have beauty, originality, or artistic merit have no traction because they miss the operational logic of the system. They repeat the habits of *l'art pour l'art*, but in reverse: critique for critique's sake. "AI slop" may well be ugly or banal—but to fixate on style, fidelity, or taste is to conflate symptom with system. The task is not to look harder at the images, but to look away from them, toward the infrastructures that produce them, and the logics they obey. Neural exchange value, as we have described it, is not simply a property of embeddings; it is a mode of organizing and simulating mental (and, in some cases, physical) labor.

When every modality is commensurable with every other—when words can become images, sounds can become captions, concepts can become prompts—this leads, almost inevitably, to media collapse. Generative AI completes what digital media began but could never quite achieve: the integration of all media into a single field of exchange. Neural exchange value is just one name for this process. There are others still to be formulated. But all belong, in one way or another, to the wider political economy of machine learning.

The next step is to make that economy explicit. This means reconnecting the study of artificial intelligence to ongoing critiques of labor, colonialism, and automation; it means treating vector spaces not only as technical objects, but as zones of work. And it means drawing on the growing body of theory (e.g., Pasquinelli 2023; Gray and Suri 2019) that insists on labor as the substrate of intelligent media.

This brings us back, full circle, to Phil Agre's revolutionary insight: that no analysis of intelligent systems can afford to separate technical development, philosophical reasoning, and materialist critique. Each must be entangled with the others, because the systems themselves are entangled. A materialist study of artificial intelligence must mirror the object it seeks to understand. It must think in three registers at once. Only then might critique become adequate to the task.

Notes

1 *Capital* is, of course, full of analogies to the hard sciences; in the foreword to the German edition, Marx explains that the value form of the commodity is analogous to the cell in biology.

2 A slightly parallel but also more metaphorical reading has recently been proposed by Mikael Brunila (2025), who understands the ideological undercurrent of embedding spaces as "cosine capital."

3 This is not too far from the case of networks trained on ImageNet-1k (whose one thousand categories are largely either consumer goods or animal species), which are broadly a standard for image embeddings.

4 Previously, each layer had to be trained through poor local approximations: for instance, Soviet-Ukrainian mathematician Alexey Grigorevich Ivakhnenko's deep networks of the mid-1960s were trained using layer-by-layer least-squares fitting.

5 For instance, in Hinton (1986), twenty-four input units are passed through six hidden units, the internal representations. As we have sketched out in the previous chapter, Hinton was, of course, not alone in proposing the bottleneck (or neural compression in general) as a useful way to learn internal representations; but he was a major protagonist—perhaps the major protagonist (along with Barlow)—in its development.

6 Classic cases of ambiguity include bistable illusions such as the duck-rabbit (made famous by Wittgenstein) or the Necker cube, but, in truth, the problem is far more widespread: Practically every real-world visual scene will include some sort of ambiguity.

7 We are not quite claiming here that Hinton's PhD work was the ground zero of neural embeddings—though it was an important early version of a probabilistic internal representation in computer vision. It is nonetheless remarkable that Hinton, who would go on to very much define the field of neural network–based computer vision and AI, saw the analogy quite clearly already in the late 1970s.

8 Hinton's work on embeddings is also closely connected to dimensionality reduction, as we explored in the previous chapter; Hinton is also one of the authors of perhaps the most commonly-used dimensionality reduction algorithm, t-SNE.

9 See Gastaldi (2021) for an excellent account of the entanglement of structuralism and embedding, from Ferdinand de Saussure to word2vec. However, Gastaldi's main argument is that there is still a structuralist theory of language encoded in the embedding space. We argue that this is not the case. The embedding strives to shed any theory of media, text, or pictures—Saussure as well as David Marr—and, at least in Hinton's formulation, attempts to explicitly transcend structure. Whether we actually buy into this transcendence is another matter—but it is important to acknowledge it precisely to be able to critique it. On the general relationship between structuralism, cybernetics, and digital media, see also the work of Bernard Dionysius Geoghegan (2022).

10 This represents another form of commensurability, one that is also historical—AI models function as snapshots of culture, akin to a kind of "snapshot ecology." This resonates with Searle's argument, which hinges on the separation of form and content, syntax and semantics. However, this very distinction is troubled by contemporary AI architectures, where meaning emerges not through fixed symbolic manipulation, but through the fluid dynamics of embeddings. A related thread can be found in Gilbert Simondon's notion of individuation, which challenges rigid categorizations in favor of continuous processes of becoming. Furthermore, Edmund Husserl's discussion of geometry as an idealized system of meaning—taken up by Derrida in *Edmund Husserl's Origin of Geometry* (1989)—anticipates these tensions between structure and transformation in computational models.

11 On the index, indexicality, and AI, see Weatherby and Justie (2022).

12 As discussed in the previous chapter. For example, in the paper presenting the *Neocognitron*: "A neural network model for a mechanism of visual pattern recognition is proposed in this paper. [. . .] The network has been *simulated* on a digital computer" (Fukushima 1980, 193, emphasis ours).

13 In fact, the situation is even more complex. There are results in model compression research, such as Sarah Hooker and colleagues' (2020) seminal work, which show that violently increasing quantization has an implication on model bias; but that dynamic is quite different to the target of our analysis.

14 Kittler's argument is, of course, not driven by naïveté, but by ideology critique. His idea is not to abandon the critique of the "soft" upper levels of the stack, but instead expand Foucauldian discursive analysis such that it would reach all the way down the stack—a "Silicon Sociology," as Geoffrey Winthrop-Young (2000) has called it—and include the ideologies embedded in the hardware—which, in the last instance, turns out to be "hardware for the destruction of other hardware," that is, war technology.

15 For a discussion of "latent spaces," or more generally embedding spaces, in architectures such as Stable Diffusion, which seem not to naturally permit such a concept, see Schaerf et al. (2025).

16 A not-quite-right translation of *Aufgabe*, which means "task" *and* "to give up." The discussion becomes almost circular when word embeddings are positioned as a potential "solution" to the problem Benjamin brings up, such as in Dougherty (2025).

17 See especially Paolo Caffoni's (2025) work on neural machine translation. For earlier work on linguistic and algorithmic capital, see Kaplan (2014) and especially Pasquinelli (2009).

18 German filmmaker Harun Farocki—in the words of fellow artist Trevor Paglen (2014, 1)—was "one of the first to notice that image-making machines and algorithms were poised to inaugurate a new visual regime. Instead of simply representing things in the world, the machines and their images were starting to 'do' things in the world."

19 Foundation models (which claim to be deployable in any context, rather than being specialized on a single task such as medical imaging) and multimodal

models (which can deal with multiple modalities of data) are in this sense expressions of the same totalizing ambition.

20 See Crawford and Paglen (2021).

21 Including transformer-based large language models, variational autoencoders, and even simple Gaussian mixture models.

References

Adorno, Theodor W., and Max Horkheimer. 2016. *Dialectic of Enlightenment*. Translated by John C. Cumming. Verso.

Amoore, Louise. 2024. "The Deep Border." *Political Geography* 109 (March). https://doi.org/10.1016/j.polgeo.2021.102547.

Baxandall, Michael. 1979. "The Language of Art History." *New Literary History* 10, no. 3: 453–65.

Benjamin, Walter. 2002. "The Task of the Translator." In *Selected Writings*, edited by Marcus Bullock and Michael W. Jennings. Harvard University Press.

Bennett, Karen. 2022. "Systems of Exchange: Translation, Money, and the Ecological Turn." *Translation Matters* 4, no. 2: 6–22.

Berardi, Franco Bifo. 2012. *The Uprising: On Poetry and Finance.* MIT Press.

Bolz, Norbert, Friedrich A. Kittler, and Christoph Tholen, eds. 1994. *Computer als Medium*. Fink.

Bowker, Geoffrey C., and Susan L. Star. 2000. *Sorting Things Out: Classification and Its Consequences.* MIT Press.

Brunila, Mikael. 2025. "Taking AI into the Tunnels." *e-flux*, no. 151 (February). https://www.e-flux.com/journal/151/652643/taking-ai-into-the-tunnels.

Caffoni, Paolo. 2025. "The Resource Debate in Machine Translation and Large Language Models." In *Handbuch Soziale Praktiken und Digitale Alltagswelten*, edited by Christian Pentzold, Bente Schöning, and Tobias Röhl. Springer VS.

Conneau, Alexis, Holger Schwenk, Loïc Barrault, and Yann LeCun. 2017. "Very Deep Convolutional Networks for Text Classification." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, edited by Mirella Lapata, Phil Blunsom, and Alexander Koller. Association for Computational Linguistics.

Crawford, Kate, and Trevor Paglen. 2021. "Excavating AI: The Politics of Images in Machine Learning Training Sets." *AI and Society* 36, no. 4: 1105–16.

Derrida, Jacques. 1989. *Edmund Husserl's Origin of Geometry: An Introduction*. Translated by John P. Leavey Jr. University of Nebraska Press.

Devlin, Hannah. 2015. "Google a Step Closer to Developing Machines with Human-Like Intelligence." *Guardian*, May 21. https://www.theguardian.com/science/2015/may/21/google-a-step-closer-to-developing-machines-with-human-like-intelligence.

Dick, Stephanie. 2015. "Of Models and Machines: Implementing Bounded Rationality." *Isis* 106, no. 3: 623–34.

Donahue, Chris, Julian McAuley, and Miller Puckette. 2018. "Adversarial Audio Synthesis." In *International Conference on Learning Representations*. OpenReview.

Dougherty, Stephen. 2025. “From Pure Language to Word Embeddings: Benjamin, Metaphysics, and Machine Translation.” *Translation Review* 121, no. 1: 15–24.

Fletcher, Angus. 2021. “Why Computers Will Never Read (or Write) Literature: A Logical Proof and a Narrative.” *Narrative* 29, no. 1: 1–28.

Fukushima, Kunihiko. 1980. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position.” *Biological Cybernetics* 36, no. 4: 193–202.

Galloway, Alexander R., and Jason R. LaRivière. 2017. “Compression in Philosophy.” *boundary 2* 44, no. 1 (2017): 125–47.

Gastaldi, Juan Luis. 2021. “Why Can Computers Understand Natural Language? The Structuralist Image of Language Behind Word Embeddings.” *Philosophy and Technology* 34, no. 1: 149–214.

Geoghegan, Bernard Dionysius. 2022. *Code: From Information Theory to French Theory.* Duke University Press.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, et al. 2023. “Imagebind: One Embedding Space to Bind Them All.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society.

Graeber, David. 2001. *Toward an Anthropological Theory of Value: The False Coin of Our Own Dreams.* Palgrave.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Harper Business.

Hao, Shibo, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, et al. 2024. “Training Large Language Models to Reason in a Continuous Latent Space.” Preprint, arXiv, December 9, 2024, 2412.06769.

Hinton, Geoffrey E. 1977. “Relaxation and Its Role in Vision.” PhD diss., University of Edinburgh.

Hinton, Geoffrey E. 1984. “Distributed Representations.” Technical Report CMU-CS-84-157. Carnegie Mellon University, Department of Computer Science.

Hinton, Geoffrey E. 1986. “Learning Distributed Representations of Concepts.” In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates.

Hinton, Geoffrey E. 1990. “Mapping Part-Whole Hierarchies into Connectionist Networks.” *Artificial Intelligence* 46, nos. 1–2: 47–75.

Hinton, Geoffrey E. 2020. Reddit post. April 23. https://www.reddit.com/r/MachineLearning/comments/g5ali0/comment/fo8rew9/?.

Hinton, Geoffrey E., and Zoubin Ghahramani. 1997. “Generative Models for Discovering Sparse Distributed Representations.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 352, no. 1358: 1177–90.

Hooker, Sara, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. “Characterising Bias in Compressed Models.” Preprint, arXiv, last revised December 18, 2020, 2010.03058.

Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. “Position: The Platonic Representation Hypothesis.” In *Proceedings of the 41st International Conference on Machine Learning*, edited by Samory Kpotufe, Pritish Kamath, and Huishuai Zhang. PMLR.

Impett, Leonardo, and Fabian Offert. 2024. "There Is a Digital Art History." *Visual Resources* 38, no. 2: 186–209.

Joque, Justin. 2022. *Revolutionary Mathematics: Artificial Intelligence, Statistics, and the Logic of Capitalism.* Verso.

Kaplan, Frederic. 2014. "Linguistic Capitalism and Algorithmic Mediation." *Representations* 127, no. 1: 57–63.

Kittler, Friedrich. 2012. "There Is No Software." In *Literature, Media, Information Systems.* Routledge.

LeCun, Yann. 2016. "Deep Learning and the Future of AI." Lecture presented at the CERN Colloquium, CERN, Geneva, March 24, 2016. https://indico.cern.ch/event/510372/.

Liang, Victor Weixin, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. 2022. "Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning." *Advances in Neural Information Processing Systems* 35: 17612–25.

Macherey, Pierre. 2016. "On the Process of Exposition of *Capital* (The Work of Concepts)." In *Reading Capital: The Complete Edition.* Edited by Louis Althusser, Étienne Balibar, Roger Establet, Pierre Macherey, and Jacques Rancière. Verso.

Mackenzie, Adrian. 2015. "The Production of Prediction: What Does Machine Learning Want?" *European Journal of Cultural Studies* 18, no. 4–5: 429–45.

Mackenzie, Adrian. 2017. *Machine Learners: Archaeology of a Data Practice*. MIT Press.

Marx, Karl. 1976. *Capital: A Critique of Political Economy, Volume One.* Translated by Ben Fowkes. Penguin.

Marx, Karl. 2005. *Grundrisse: Foundations of the Critique of Political Economy*. Penguin.

Moretti, Franco. 2013. *Operationalizing: Or, the Function of Measurement in Modern Literary Theory*. Stanford Literary Lab Pamphlet no. 6.

Offert, Fabian. 2023. "On the Concept of History (in Foundation Models)." *image* 37: 121–34.

Offert, Fabian, and Peter Bell. 2023. "imgs.ai: A Multimodal Search Engine for Digital Art History." *International Journal for Digital Art History* 9: 5–28.

Offert, Fabian, Paul Kim, and Qiaoyu Cai. 2024. "Synthesizing Proteins on the Graphics Card. Protein Folding and the Limits of Critical AI Studies." Preprint, arXiv, last revised December 7, 2024, 2405.09788.

Offert, Fabian, and Thao Phan. 2024. "A Sign That Spells: Machinic Concepts and the Racial Politics of Generative AI." *Journal of Digital Social Research* 6, no. 4: 49–59.

Paglen, Trevor. 2014. "Operational Images." *e-flux*, no. 59 (November): 1–3.

Pasquinelli, Matteo. 2009. "L'algoritmo PageRank di Google: Diagramma del capitalismo cognitivo e rentier dell'intelletto comune." *Sociologia del lavoro* 115: 1–11.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence.* Verso.

Phan, Thao, and Scott Wark. 2021. "Racial Formations as Data Formations." *Big Data and Society* 8, no. 2. https://doi.org/10.1177/20539517211046377.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. 2021. "Learning Transferable Visual Models from Natural Language

Supervision." In *Proceedings of the 38th International Conference on Machine Learning*, edited by Marina Meila and Tong Zhang. PMLR.

Rossi-Landi, Ferruccio. 1992. *Il linguaggio come lavoro e come mercato: Una teoria della produzione e scambio linguistico.* Bompiani.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323, no. 6088: 533–36.

Schaerf, Ludovica, Andrea Alfarano, Fabrizio Silvestri, and Leonardo Impett. 2025. "Training-Free Style and Content Transfer by Leveraging U-Net Skip Connections in Stable Diffusion." Preprint, arXiv, January 24, 2025, 2501.14524v1.

Shmilovici, Armin, Yoav Kahiri, Irad Ben-Gal, and Shmuel Hauser. 2009. "Measuring the Efficiency of the Intraday Forex Market with a Universal Data Compression Algorithm." *Computational Economics* 33: 131–54.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. "AI Models Collapse When Trained on Recursively Generated Data." *Nature* 631, no. 8022: 755–59.

Sohn-Rethel, Alfred. 1970. *Geistige und körperliche Arbeit: Zur Theorie gesellschaftlicher Synthesis.* Suhrkamp.

Sprinker, Michael, ed. 2008. *Ghostly Demarcations: A Symposium on Jacques Derrida's Specters of Marx.* Verso.

Srnicek, Nick. 2017. *Platform Capitalism.* Polity.

Underwood, Ted. 2021. "Mapping the Latent Spaces of Culture." *The Stone and the Shell*, October 21. https://tedunderwood.com/2021/10/21/latent-spaces-of-culture/.

Venugopal, Ashish, Andreas Zollmann, Franz Och, John Blatz, and Jeff Uszkoreit. 2011. "Watermarking the Outputs of Structured Prediction with an Application in Statistical Machine Translation." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Weatherby, Leif, and Brian Justie. 2022. "Indexical AI." *Critical Inquiry* 48, no. 2: 381–415.

Winthrop-Young, Geoffrey. 2000. "Silicon Sociology, or, Two Kings on Hegel's Throne? Kittler, Luhmann, and the Posthuman Merger of German Media Theory." *Yale Journal of Criticism* 13, no. 2: 391–420.

Wu, Ledell, Adam Fisch, Sumit Chopra, et al. 2018. "StarSpace: Embed All the Things!" *Proceedings of the AAAI Conference on Artificial Intelligence* 32, no. 1. https://doi.org/10.1609/aaai.v32i1.11789.

Zhang, Lingze, and Ellie Pavlick. 2025. "Does Training on Synthetic Data Make Models Less Robust?" Preprint, arXiv, February 11, 2025, 2502.07164.

Authors

Johanna Drucker is Distinguished Professor and Breslauer Professor Emerita, Department of Information Studies at UCLA. Her latest book is *Inventing the Alphabet: The Origins of Letters from Antiquity to the Present.*

Leonardo Impett is leader of the machine visual culture research group at the Bibliotheca Hertziana–Max Planck Institute for Art History, and assistant professor of digital humanities at the Faculty of English, University of Cambridge.

Fabian Offert is assistant professor of the history and theory of the digital humanities and director of the Center for the Humanities and Machine Learning at the University of California, Santa Barbara.

Media Studies

Leonardo Impett, Fabian Offert and Johanna Drucker

Vector Media

Contemporary artificial intelligence systems translate different media forms, such as text, images, and sound, into a shared vector space. *Vector Media* is the first comprehensive history and theory of this vector space. Along with introducing the reader to latent theories of image and knowledge, the book moves the debate of critical artificial intelligence studies away from generated outputs and toward internal mechanisms as the decisive sites where cultural logics are encoded.

Inspired by Phil Agre's critical technical practice, *Vector Media* proposes a new way to theorize artificial intelligence systems as epistemic objects whose politics are determined by their logic of representation.

Published by the
University of Minnesota Press
in collaboration with meson press

ISBN 978-1-5179-2167-5

9 781517 921675