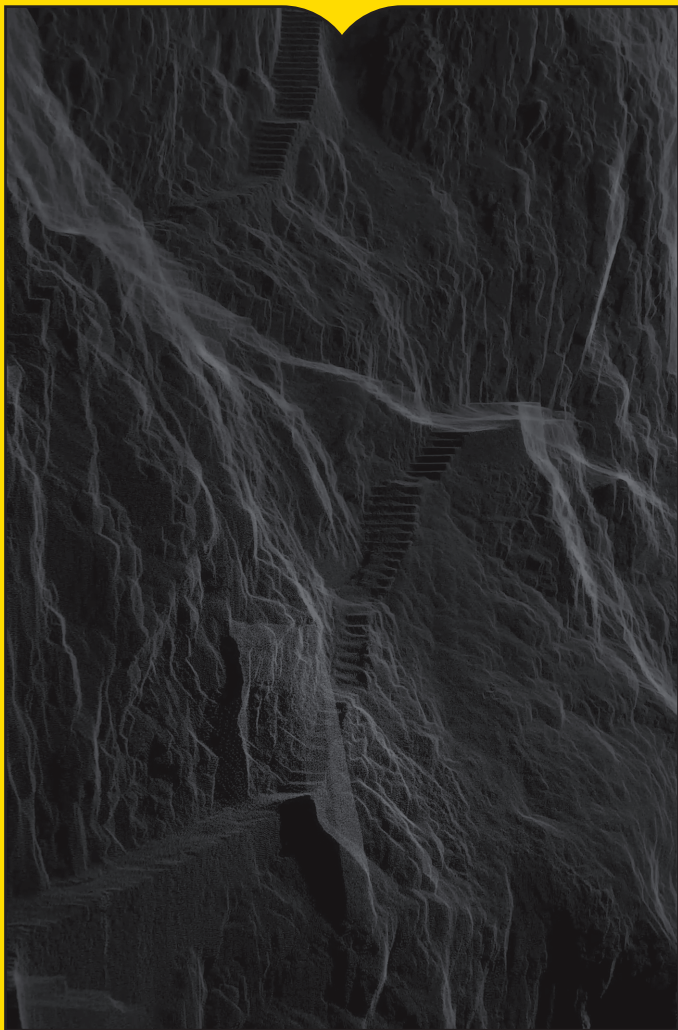


M. Beatrice Fazi
Cave Machines: On the Platonic
Representation Hypothesis



M. Beatrice Fazi
Cave Machines:
On the Platonic Representation Hypothesis

Cave

The story is old, and it goes like this. In an underground cave, a group of prisoners has been chained since childhood. They are bound by their necks and legs, forced to face the cave wall, and unable to turn their heads. Behind them, a fire burns. Between the fire and the prisoners runs a raised walkway, and along it a parapet, like a puppeteer's screen. People walk this path, raising objects above the parapet – human and animal figures carved in stone and wood. The parapet conceals the carriers; only what they hold clears its edge. The firelight throws the shadows of these objects, and only these objects, onto the wall the prisoners face. Some of the carriers speak as they pass; their voices bounce off that same wall, and the prisoners, hearing the sound come from the shadows, conclude that it is the shadows who talk. Because these shadows are all they have ever seen and heard, they take them for reality itself. They hold competitions – honouring whoever can name a shadow most surely, or best guess which will appear next – and praise one another for their intelligence.

One day, a prisoner is freed. At first, freedom is pain. He is made to stand and turn. The fire blinds him. The figures he now sees appear less real than the shadows he has always known. He wants to turn back; he is told to keep going. He is hauled up the long slope, out of the mouth of the cave, into the open air. The sun is unbearable. He covers his eyes. Gradually, he learns to look: first at shadows on

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

165.21:1Platon
141.131:004

FAZI, M. Beatrice

Cave machines : on the platonic representation hypothesis / M. Beatrice Fazi. -
Ljubljana : Aksioma - Institute for Contemporary Art ; Vienna : University of Applied
Arts, 2026. - (PostScriptUM ; #56)

ISBN 978-961-7173-70-3 (Aksioma)
COBISS.SI-ID 291591171

the ground, then at reflections in water, then at the things that cast them. He studies the moon and stars; last of all, the sun itself. He understands what the sun is: it makes the seasons and the years, it governs everything in the visible world, and it is, in some way, the cause of all that he and the others ever saw below. Everything he once took to be real was only a faint, distorted copy.

Here the story turns, because the prisoner remembers the others, and pities them, and goes back down. But his eyes, accustomed now to brightness, can no longer make out the shadows that everyone else still treats as truth. He stumbles. He fumbles the naming games he would once have won. The other prisoners mock him. His journey outside has ruined him, they conclude; leaving the cave is a terrible idea. Better not to climb at all. And if anyone ever came to lead them up, they would fight back. If they could, they would kill him.

This tale is, of course, Plato's *allegory of the cave*, from book seven of the *Republic* – one of the oldest thought experiments in philosophy and, as such, a vehicle: it asks its audience to imagine a world outside the one they know.¹ It has been doing that work for twenty-four hundred years, and in doing so has become the founding scene of Western representationalism – not a theory of mind but one of copies and originals. Its axioms are three. Behind the images there is a reality the images copy; the copies are ranked by fidelity to it; and truth lies off-screen, outside, reachable only by turning away from the images towards the things themselves. This grammar of surfaces and depths has left later

philosophy two options: to elaborate it, or to try to escape it. Either way, *representation* comes off badly. Elaborate the Platonic grammar and representation is a deficient copy by definition. Escape it and representation fares no better: the exit abolishes rather than redeems it. Within this legacy, to call something a representation is already to demote it – to say that it is not the thing but its stand-in, the deferral of a presence it can never deliver. Much of the subsequent history of thought is an attempt either to overturn this verdict or to live with it (“Wake up, Neo”).²

The allegory's latest afterlife is also its least likely. In 2024, four researchers at MIT gave Plato's name to an empirical hypothesis about machines – a conjecture argued, pointedly, as a position paper rather than reported as a result: that artificial neural networks, trained on different data and for different tasks, are converging on a single shared statistical model of the reality that produced that data, a single ground truth of which each model's representation is a better or worse approximation.³ As vision and language models grow larger, Huh, Cheung, Wang and Isola argue, these models' internal representations align – across architectures and across modalities, models come to agree, more and more, on the distances between data points.

The parallel with Plato's allegory of the cave is the authors' own. There is a world, and the data drawn from it to feed the machine are projections of that world – its shadows. The machine, trained on those shadows, assembles

1 Plato. (2012). *The Republic* (C. J. Rowe, Trans.). Penguin.

2 Wachowski, L., & Wachowski, L. (Directors). (1999). *The Matrix* [Film]. Warner Bros.

3 Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic Representation Hypothesis. *Proceedings of the 41st International Conference on Machine Learning*, 235, 20617–20642.

a representation that reaches back towards whatever cast them. An image of a dog chasing a ball and the sentence “a dog chases a ball” are two shadows of one event, cast onto different walls. Each model is a prisoner before a single wall: the vision model sees only pixels, the language model only tokens. The prisoners do not even play the same game – some are trained to name the shadows, some to fill in a shadow that has been masked, some to guess which comes next. The hypothesis the MIT researchers propose is that any of these games, played well enough, ends in the same place: each prisoner, in learning its own wall, is forced to model what casts onto it, and so prisoners who have never met come to agree. What they converge towards, the authors suggest, is a representation of the world that generates every projection, of which each actual model is a better or worse approximation. Hence the name they give it: the *Platonic Representation Hypothesis* (PRH), after the cave and after Plato’s ideal reality behind appearances. The cave’s metaphysics has been recast as a falsifiable claim about embedding spaces.

Convergence

Media studies have been here before, and the precedent is cautionary. Once the media were many. The printed page, the telephone wire, the broadcast tower: *convergence* named their collapse into one – a “convergence of modes” that came to pass much as foretold,⁴ the partitions that had kept post from telephone and telephone from television dissolving as everything learned to travel as signal. A companion

prophecy was borne out with it: the airwaves and the wires would trade places, broadcast moving to the cable, the telephone to the air. Negroponte pressed the underlying point into a slogan at the MIT Media Lab: whatever can become a bit will, and once it is a bit it is the same stuff as every other bit.⁵ Kittler put it most starkly: digitisation erases the differences between media altogether; sound, image and writing, once irreducibly distinct, are levelled into number, and the concept of the medium dissolves with them (“a total media link on a digital base”).⁶ Manovich specified the operation: numerical representation beneath, transcoding across.⁷

All of this is convergence downwards: media meeting in the common code beneath them, where difference is flattened into the same. The argument is that digitality collapses format-specific infrastructures into one. The PRH, however, proposes convergence in the other direction, upwards. The models do not meet in their medium; they meet past it. They converge not on the vectors they are made of but on the statistical representation of a world these vectors are already meant to be representations of. In other words, what converges is the alignment of those already representational bearers.

Both in media and in models, the concept of convergence travels carrying assumptions it never fully earned. Theses about convergence convince because they are half true, and the half that is true lends its credit to the half that is not. The media case shows how. There the convergence

5 Negroponte, N. (1995). *Being Digital*. Knopf.

6 Kittler, F. A. (1999). *Gramophone, Film, Typewriter* (G. Winthrop-Young & M. Wutz, Trans.). Stanford University Press, p. 2.

7 Manovich, L. (2001). *The Language of New Media*. MIT Press.

4 de Sola Pool, I. (1983). *Technologies of Freedom*. Harvard University Press, p. 23.

[...]

M. Beatrice Fazi
Cave Machines:
On the Platonic Representation Hypothesis

PostScriptUM #56
Series edited by Janez Fakin Janša
Print on demand



Publisher: Aksioma – Institute for Contemporary Art, Ljubljana
www.aksioma.org | aksioma@aksioma.org
Represented by: Marcela Okretič

di:'angewandte

Universität für angewandte Kunst Wien
University of Applied Arts Vienna

Co-publisher: University of Applied Arts Vienna
www.dieangewandte.at/en | info@uni-ak.ac.at

Issue edited by Marko Bauer

Proofreading: Miha Šuštar
Design: Federico Antonini
Layout: Oskar Kandare

Cover image: Swamp_Matter (Eva Garibaldi, Ana Laura Richter),
still from the video *The Stones Only Appear To Be Non-Living*, 2024

© Aksioma | All text and image rights reserved by the author
Ljubljana, October 2026

Supported by the Ministry of Culture of the Republic of Slovenia
and the Municipality of Ljubljana.



City of
Ljubljana



REPUBLIC OF SLOVENIA
MINISTRY OF CULTURE



Weibel
Institute
for Digital
Cultures