Aesthetica Edizioni

Preprint
Periodico quadrimestrale
in collaborazione con la Società Italiana di Estetica

N. 131
gennaio-aprile 2026

Automating Sensibility

a cura di

Luca Capone, Felice Cimatti

Aesthetica Edizioni

2026 Aesthetica Edizioni

E-ISSN 2785-4442
ISBN 9788877262950

www.aestheticaedizioni.it
info@aestheticaedizioni.it

Sommario

Introduzione 7
Automating sensibility: contributions on the relationship between human aesthesis and artificial intelligence models
Luca Capone, Felice Cimatti

Seven Arguments about AI Images and Generative Media 13
Lev Manovich

Sulla natura dell'immagine:
Language Model e Visual Language Model 37
Luca Capone, Felice Cimatti

Reimagining AI Perception: Cybernetics, Extractivism, and the DIY Turn 55
Samuele Sartori

What Effect Does Your Nonhuman Touch Have on Me? An Aesthetic Approach to Human-Robot Interaction 71
Chiara Cappelletto, Giulio Galimberti

Re-examining traditional concepts in aesthetics from the lens of AI 89
Nicola Di Stefano

Di cosa parliamo quando parliamo di creatività dell'IA? 105
Prolegomeni a una teoria dell'arte post-umana
Dario Cecchi

What AI philosophers lack 123
Simona Chiodo

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221511

Introduzione
Automating sensibility: contributions on the relationship between human aesthesis and artificial intelligence models

Luca Capone, Felice Cimatti

Research in the field of artificial intelligence (AI) has predominantly focused on inquiries into human cognition, comprehension, rationality, and the potential for automating these abilities through training or coding. Recent advancements in natural language processing have been the subject of extensive literature, confirming this trend. Running parallel to this line of research, a related discipline that might aptly be termed *artificial sensibility* has emerged. Models that generate images from textual prompts, compose musical themes, support diagnosis through medical imaging, or detect solemn or ironic nuances in text, all fall within the purview of this research domain.

Despite the possibility of framing these models' activities from such a perspective, there is a notable absence of substantial literature dedicated to contextualizing the issue of aesthetics. Frequently, even Computer Vision models are incorporated into the broader AI discourse, where the aesthetic dimension is subsumed under the intellectual one, thus relegating sensibility to a subordinate role. This points to a broader issue: with respect to sensibility and the aesthetic dimension, AI still struggles to constitute itself as a clearly delimited object of inquiry. AI, in other words, has entered aesthetics less as a single object than as a multiplicity of functions, objects, and conditions: a) generative models introduce new representational entities into the world, objects that may (or may not) acquire aesthetic relevance; b) Large Language Models can themselves perform aesthetic functions, detecting and describing aesthetic features of representations (graphic, textual, audiovisual, etc.); c) infrastructural and cultural shifts driven by AI reconfigure the conditions under which aesthetic objects appear. They alter how representations are produced, circulated, interpreted, and valued, while also reshaping what "making" can mean when creation is increasingly mediated by prediction, automation, and large-scale infrastructural extraction.

Across these domains, the most persistent questions are not merely "Can AI create?" but: What kind of perception does AI enact? What kind of image is an AI image? How are agency, authorship, and responsibility distributed across systems, designers, datasets, institutions, and audiences? And, more importantly, what do these shifts imply for human sensibility, agency, authorship, and more broadly for practices of sense-making? This issue of Aesthetica Preprint brings together seven contributions (in Italian and English) that collectively map AI's aesthetic problem-space along three complementary sections: (1) generative models across the history of representation and mediation; (2) the epistemologies and embodiments of machine perception, from cybernetic lineages to tactile encounters; (3) creativity, authorship, and the philosophical limits of automation.

Generative models across the history of representation and mediation (Manovich; Capone & Cimatti)

The first section situates today's "AI images" within a longer genealogy of representational techniques and theories. Contemporary generative systems are often framed as a rupture; yet ruptures become legible only when they are placed in relation to earlier reorganizations of seeing and making, such as linear perspective, photography, or the successive waves of digital tools that progressively automated parts of visual production. Within this frame, the novelty of generative media is not simply that images are produced by machines, but that they are produced *predictively*: as statistically guided syntheses drawn from large repositories of existing cultural artifacts. This predictive regime compels aesthetics to recalibrate its own vocabulary (form and content, subject and style, originality and repetition) because it makes newly visible how cultural regularities (the frequent, the typical, the "average") can be recombined at scale, while the rare, the idiosyncratic, and the context-bound risk being flattened into stereotypical or idealized outputs. Lev Manovich's contribution provides the historical and conceptual scaffolding for this recalibration. By treating generative media as a media revolution comparable in scope to photography or perspective, he clarifies what is specifically "new" in AI images: not a mysterious creative faculty, but a method of cultural production based on training over vast datasets and on the automation of choices that used to be made by skilled creators (selection, variation, stylistic transformation, compositing). His arguments show how the very

category of AI functions culturally (what counts as AI shifts as soon as automation becomes familiar), and how generative systems reorganize the ecology of images, expanding the space of possible aesthetics while also intensifying pressures toward recognizable conventions and high-frequency visual patterns.

If Manovich maps the media-historical shift, Luca Capone and Felice Cimatti focus on the main philosophical problem that text-to-image and image-to-text systems force into the foreground: the relation between words and images. Starting from the apparently banal loop "prompt → generated image → linguistic description" they ask what must be the case for such translation to be possible at all. Against the common intuition that images essentially contain an ineffable excess ("the visible is not reducible to the sayable"), they propose reading that slogan not as an empirical report about an ontological remainder, but as a *grammatical* proposition: a rule that organizes our practices of meaning without deciding the nature of images as such. On this basis, the paper foregrounds a theme that resonates across the whole issue: indeterminacy is not merely a limitation of models (or of humans), but a structural feature of signification. Images and verbal expressions are both "indefinitely determinable", and what generative systems reveal is less a clean conversion between two separate realms than a shared semiotic field in which meaning is produced, stabilized, and renegotiated through use.

The epistemologies and embodiments of machine perception, from cybernetic lineages to tactile encounters (Sartori; Cappelletto & Galimberti)

The second section shifts from images as outputs to perception as operativity. Sartori argues that contemporary AI inherits a bottom-up cybernetic model of learning and behavior formation, yet this inheritance is partially occluded by an extractivist paradigm that governs data, hardware, and opacity. The result is not one black box but many: we may not understand internal model processes; we may not know the objects of digital perception (training data); and we may not even be able to locate the "body" of AI, distributed across devices and infrastructures. Against this, Sartori proposes a counter-genealogy: DIY and performance-art practices that rebuild perception as situated, relational, and materially legible, projects that replace indiscriminate harvesting with consent, participation, and a consciously designed "perceptual communication"

between human and technological bodies. The aesthetic stakes here are political as much as experiential: the form of perception a technology enacts is inseparable from the social relations and ecological costs through which it is produced.

Cappelletto and Galimberti then bring the question of perception into the register of touch. Resisting brainocentric models of human–robot interaction, they focus on the bodily dimension of HRI and especially on thermoception, a sensory modality often neglected despite its strong influence on affect, trust, and willingness to cooperate. Their experimental work with the NAO robot tests how different thermal conditions in robotic grasping modulate participants' experience and acceptance, and suggests that "a surplus of warmth," rather than simple temperature matching, can increase willingness to engage. Here AI's aesthetic import is not reducible to representation: it is a material partner that can disrupt bodily transparency and reconfigure expectations through its heterogeneous affordances.

Creativity, authorship, and the philosophical limits of automation (Di Stefano; Cecchi; Chiodo)

The third section addresses creativity and meaning, precisely where public discourse most quickly oscillates between hype ("AI is creative") and dismissal ("it's just statistics"). Di Stefano argues that AI art is philosophically productive not because it settles the question of machine creativity, but because it forces clearer distinctions between creation, generation, and interpretation. Through examples ranging from canonical art history to conceptual art, and through the problem of authorship and audience bias, he proposes that creativity is better understood as a capacity for generating meaning and significance, something institutionally and interpretively conferred, not algorithmically contained.

Cecchi's contribution intersects with Di Stefano but shifts the emphasis: the methodological mistake is to treat AI like a new style or movement and then apply our ordinary evaluative criteria as if nothing had changed. Drawing on Garroni's framework, Cecchi recasts creativity in terms of how rules are applied, revised, and made sense of, within experience, and asks what it would mean to attribute such creativity to an AI "artist." His closing hypothesis is deliberately provocative: future AI art may take the form of "controlled hallucinations," whose sense emerges after the fact through negotiation between systems and publics, an art that is post-human

not because humans disappear, but because meaning is redistributed across heterogeneous agents and interpretive practices.

Finally, Chiodo's essay pushes the issue's stakes beyond art to philosophy itself. Taking a contemporary "AI philosophy" experiment as a point of departure, she argues that even when generative systems imitate philosophical answers, they lack what human philosophers have: aesthetic existences and, therefore, meta-sensemaking, the capacity to live, feel, and reflect upon the existential weight of the questions at stake. This difference, for Chiodo, connects to a broader societal tendency: exchanging the burden of autonomy for the relief of automation, especially under conditions of acceleration, uncertainty, and performance pressure. Ending here is not pessimism; it is a clarifying demand: if AI systems increasingly mediate the production of answers (aesthetic, practical, philosophical), we need conceptual and political tools to know what we are consulting, and what kind of life, sense, and responsibility we might be trading away.

Taken together, the seven papers do not offer a single verdict on AI's aesthetic status. Instead, they build a shared map of problems: prediction and representation; sign and image; black box and participation; touch and affect; creativity and interpretation; autonomy and automation. The special issue invites readers to treat AI not as an external "theme" for aesthetics, but as a pressure test for aesthetics itself, forcing the discipline to rethink what counts as sensibility, what counts as an image, what counts as creativity, and what counts as meaning when sense is increasingly co-produced by humans and machines.

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221512

Seven Arguments about AI Images and Generative Media

Lev Manovich[*, **]

Abstract

The article provides an analysis of AI-generated images and generative media as a major new stage in the history of visual culture, considered alongside earlier media transformations such as photography, 3D computer graphics, and the web. It argues that generative AI should be understood not only as a technical innovation but also as a cultural and aesthetic system shaped by vast archives of existing media. Through seven interconnected arguments, the discussion shows how AI media extend traditions associated with modernism and database art while introducing new forms of prediction, media translation, and creative synthesis. At the same time, attention is given to the limits of these systems, particularly their tendency to privilege common patterns over rare, singular, or highly personal forms of expression. Overall, the article presents generative AI as both a continuation of earlier cultural logics and a distinct paradigm for the production and interpretation of media.

Keywords

AI, aesthetics, media, art

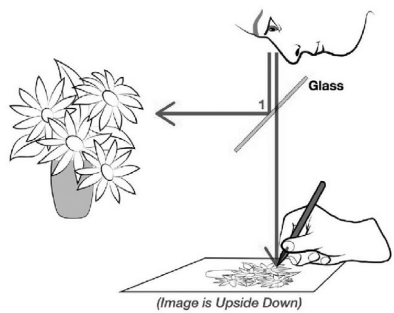


Figure 1 – Basic principle of camera lucida, optical drawing device widely used by artists and art students in the 19th century. Source: https://neolucida.com/history.

[*] lmanovich@gc.cuny.edu, City University of New York

[**] This chapter is written by me without use of any AI writing tools. In the editing stage, I use a quillbot tool that offers paraphrased versions of the selected sentences.

We appear to be in the beginning of a true revolution in media creation: the rise of generative media. I've been using computer tools for art and design since 1984, and I've seen a few major media revolutions, including the introduction of Mac computers and desktop applications for media creation and editing, the development of photorealistic 3D computer graphics and animation, the rise of the web after 1993, and the rise of social media networks after 2006. The new AI generative media revolution appears to be as significant as any of them. Indeed, it is possible that it is as significant as the invention of photography in the nineteenth century or the adoption of linear perspective in western art in the sixteenth.

If you are new to this topic, here is very brief history. Generative media revolution was in development for over 20 years. The first AI papers proposing that the vast unstructured web universe of texts, images and other cultural artifacts can be used to train computers to do various tasks appeared already in 1999-2001. In 2015 Google 'deep dream' and 'style transfer' methods attracted lots of attention: suddenly computers could create new artistic images mimicking styles of many famous artists. The release of DALL-E (Ramesh *et al.* 2021) in January 2021 was another milestone: now computers could synthesize images from text description. Midjourney, Stable Diffusion, and DALL-E 2 all contributed to the acceleration of this evolution in 2022 (Ramesh *et al.* 2022). Now synthetic images could have many aesthetics that ranges from photo-realism to any kind of physical or digital medium, including mosaics, oil paintings, street photography, or 3D CG rendering. The code for producing such images referred to as a 'model' in the field of artificial intelligence was made public in August 2022, sparking a flurry of experiments and accelerating development.

In this and the next chapters I will describe a number of characteristics of *visual generative media* in its current forms that I believe are particularly significant or novel. Some of my arguments also apply to generative media in general, but most focus on *visual media* – reflecting my own experience of using a few popular AI image tools such as Midjourney and Stable Diffusion (Rombach *et al.* 2022) (and sometimes also Runway ML) almost every day from middle of 2022 to early 2023. But first, let's define the main terms.

The Terms

In this text, 'artist' or 'creator' refers to any skilled person who creates cultural objects in any media or their combinations. The

terms 'generative media', 'AI media', 'generative AI', and 'synthetic media' are all interchangeable. They refer to the process of creating new media objects with deep neural networks, such as images, animation, video, text, music, 3D models and scenes, and other types of media. Neural networks are also used to generate specific elements and types of content, such as photorealistic human faces and human poses and movements, in addition to such objects. They can also be used in media editing, such as replacing a portion of an image or video with another content that fits spatially.

These networks are trained on vast collections of media objects already in existence. Popular artificial neural network types for media generation include diffusion models, text-to-image models, generative adversarial networks (GAN), and transformers. For the generation of still and moving images using neural networks, the terms image generation, synthetic image, AI image, and AI visuals can be used interchangeably.

Note that the word 'generative' can be also used in different ways to refer to making cultural artifacts using any algorithmic processes (as opposed to only neural networks) or even a rule-based process that does not use computers. This is how the phrases 'generative art' and 'generative design' are typically used today in cultural discourses and popular media. In this chapter I am using 'generative' in more restrictive way to designate deep network methods and apps for media generation that use these methods.

1. *'AI' as a Cultural Perception*

There is not one specific technology or a single research project called 'AI'. However, we can follow how our cultural perception of this concept evolved over time and what it was referring to in each period. In the last fifty years, when an allegedly uniquely human ability or skill is being automated by means of computer technology, we refer to it as 'AI'. Yet, as soon as this automation is seamlessly and fully successful, we tend to stop referring to it as an 'AI case'. In other words, 'AI' refers to technologies and methodologies that automate human cognitive abilities and are starting to function but aren't quite there yet. 'AI' was already present in the earliest computer media tools. The first interactive drawing and design system, Ivan Sutherland's *Sketchpad* (1961-1962), had a feature that would automatically finish any rectangles or circles you started drawing. In other words, it knew what you were trying to make. In the very broad understanding just given, this was undoubtedly 'AI' already.

My first experience with a desktop paint program running on an Apple II was in 1984, and it was truly amazing to move your mouse and see simulated paint brushstrokes appear on the screen. However, today we no longer consider this to be 'AI'. Another example would be the Photoshop function that automatically selects an outline of an object. This function was added many years ago – this, too, is 'AI' in the broad sense, yet nobody would refer to it as such today. The history of digital media systems and tools is full of such 'AI moments' – amazing at first, then taken for granted and forgotten as 'AI' after a while. (In AI history books, this phenomenon is referred to as the 'AI effect'). At the moment, 'creative AI' refers only to recently developed methods where computers transform some inputs into new media outputs (e.g., text-to-image models) and specific techniques (e.g., certain types of deep neural networks). However, we must remember that these methods are neither the first nor the last in the long history and future of simulating human art abilities or assisting humans in media creation.

2. *'Make it New': AI and Modernism*

After training on trillions of text pages or billions of art and photographic pictures taken from the web, neural networks can generate fresh texts and visuals on the level of highly competent professional writers, artists, photographers, or illustrators. These capacities of the AI systems nets are distributed over trillions of connections between billions of artificial neurons rather than determined by standard algorithms. In other words, we developed a technology that, in terms of complexity, is extremely similar to the human brain. We don't fully grasp how our AI technology works, just as we don't fully comprehend human intellect and creativity.

The current generation of generative AI systems, such as GPT and Stable Diffusion, have been trained on very large and diverse datasets consisting from billions or even trillions of individual texts, or image and text pairs. It is, however, equally interesting to limit the training data set to a specific area within the larger space of human cultural history, or to a specific set of artists from a specific historical period. Unsupervised[1] by Refik Anadol Studio (2022) is a AI art project that exemplifies these possibilities. The project uses neural networks trained on the image dataset of tens of thousands of artworks from the MoMA collection. This collection, in

[1] https://refikanadol.com/works/unsupervised/

my opinion, is one of the best representations of the most creative and experimental period in human visual history – hundred years of modern art (1870 – 1970) – as well as many important examples of artistic explorations in the subsequent decades. It captures modernist artists' feverish and relentless experiments to create new visual and communication languages and 'make it new'.

Figure 2 – Unsupervised, Refik Anadol Studio (2022).
Selected frames from the animation.

On the surface, *the logic of modernism appears to be diametrically opposed to the process of training generative AI systems*. Modern artists desired to depart from classical art and its defining characteristics such as visual symmetry, hierarchical compositions, and narrative content. In other words, their art was founded on a fundamental rejection of everything that had come before it (at least in theory, as expressed in their manifestos). Neural networks are trained in the opposite manner, by learning from historical culture and art created up to now. A neural network is analogous to a very conservative artist studying in the 'meta' 'museum without walls' that houses historical art.

But we all know that art theory and art practice are not the same thing. Modern artists did not completely reject the past and everything that came before them. Instead, *modern art developed by reinterpreting and copying images and forms from old art traditions*, such as Japanese prints (van Gogh), African sculpture

(Picasso), and Russian icons (Malevich). Thus, the artists only rejected the dominant high art paradigms of the time, realistic and salon art, but not the rest of human art history. In other words, it was deeply historicist: rather than inventing everything from scratch, it innovated by adapting certain older aesthetics to contemporary art contexts. (In the case of geometric abstract art created in 1910s, these artists used images that were already widely used in experimental psychology to study human visual sensation and perception. For the detailed analysis of these relations between modern art and experimental psychology, see Paul Vitz and Arnold Glimcher, Modern art and Modern Science: The Parallel Analysis of Vision, 1985).

When it comes to artistic AI, we should not be blinded by how these systems are trained. Yes, artificial neural networks are trained on previously created human art and culture artifacts. However, their newly generated outputs are not mechanical replicas or simulations of what has already been created. In my opinion, these are frequently *genuinely new* cultural artifacts with *previously unseen content, aesthetics, or styles*.

Of course, simply being novel does not automatically make something culturally or socially interesting or significant. Indeed, many definitions of 'creativity' agree on this point: it is the creation of something that is both original and worthwhile or useful.

However, estimating what percentage of all novel artifacts produced by generative AI are also useful and/or meaningful for a larger culture is not a feasible project at this time. For one thing, I am not aware of any systematic effort to use such systems to 'fill in', so to speak, a massive matrix of all content and aesthetic possibilities by providing millions of specifically designed prompts. Instead, it is likely that, as in every other area of popular culture, only a small number of possibilities are realized over and over by millions of users, leaving a long tail of other possibilities unrealized. So, if only a tiny fraction of the vast universe of potential AI artifacts is being realized in practice, we can't make broad statements about the originality or utility of the rest of the universe.

3. *Generative Media and Database Art*

Some AI artists such Anna Ridler[2], Sarah Meyohas[3] and Refik Anadol utilized in their works nets trained on specific datasets.

[2] https://annaridler.com/
[3] https://aiartists.org/sarah-meyohas

Many other artists, designers, architects, and technologists use networks released by other companies or research institutions that were already trained on very large datasets (e.g, Stable Diffusion), and then fine-tune them on their own data.

For example, artist Lev Pereulkov[4] fine-tuned the Stable Diffusion model 2.1 using 40 paintings by well-known 'non-conformist' artists who worked in USSR starting in the 1960s (Erik Bulatov, Ilya Kabakov, etc). Pereulkov's image series Artificial Experiments 1-10 (2023) created with this custom net, is an original piece of art that captures the artistic characteristics of these artists as well as their unique surreal and ludicrous semantics without repeating closely any of their existing works. Instead, their 'DNAs' captured by the net enable new meanings and visual concepts.

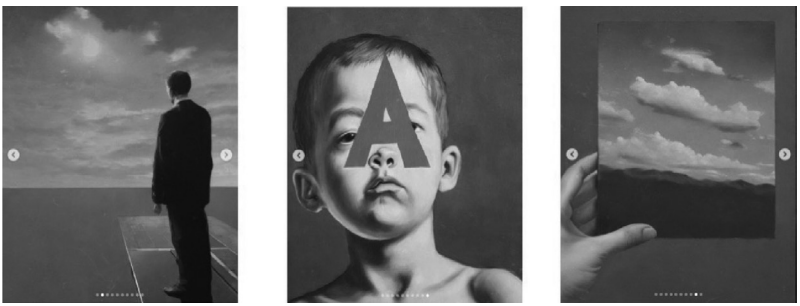


Figure 3 – Lev Pereulkov, Artificial Experiments 1-10, 2022.
Three images from the series of 10 shared on Instagram

Most of the millions of everyday people and creative professionals who employ generative media tools use them as is, and don't fine them further. This may change in the future as the techniques networks using our own data may become easier to use. But regardless of these specifics, all newly created cultural artifacts produced by trained nets have a common logic.

Unlike traditional drawings, sculptures, and paintings, generative media artifacts are not created from scratch. They are also not the result of capturing some sort of sensory phenomenon, such as photos, videos, or sound recordings. They are instead built from a large archive of other media artifacts. This generative mechanism links generative media to earlier art genres and processes.

[4] https://www.instagram.com/pereulye/

We can compare it to film editing, which first appears around 1898, or even earlier composite photography, which was popular in the nineteenth century. We can also consider specific artworks that are especially relevant, such as experimental collage film A Movie[5] (Bruce Conner, 1958) or many Nam June Park installations that feature edited fragments of TV footage.

Seeing projects like *Unsupervised* or *artificial experiments 1-10* in the context of this media creation method and its historical variations will help us understand this and many other AI artworks as art objects engaged in dialogues with art from the past, rather than as purely technological novelties or works of entertainment.

I see many relevant moments and periods when I scan the history of art, visual culture, and media for other prominent uses of this procedure. They are relevant to the current generative media not only because artists working at these times used the procedure, but also because the reason for this use was consistent in all cases. *A new accumulation and accessibility of masses of cultural artifacts led artists to create new forms of art driven from these accumulations*. Let me describe a few of these examples.

Net and digital artists created a number of works in the late 1990s and early 2000s in response to the new and rapidly expanding universe of the world wide web. Health Bunting's _readme[6] (1998), for example, is a web page containing the text of an article about the artist, with each word linked to an existing web domain corresponding to that word. Mark Napier's Shredder 1.0[7] (also 1998) presents a dynamic montage of elements that comprise numerous websites – images, texts, HTML code, and links.

Going further back in time, we find a broad cultural paradigm that was also a reaction to the accumulation of historical art and culture artifacts in easily accessible media collections. This is paradigm is known as 'post-modernism'. Post-modern artists and designers frequently used bricolage and created works consisting of quotations and references to art from the past, rejecting modernism's focus on novelty and breaking with the past.

While there are many possible explanations for the emergence of the post-modern paradigm in the 1960s and 1980s, one is relevant to our discussion. The accumulation of earlier art and media artifacts in structured and accessible collections such as slides libraries, film archives, art history textbooks with many photos of the artworks, and other formats – where different historical peri-

[5] https://en.wikipedia.org/wiki/A_Movie
[6] http://www.medienkunstnetz.de/works/readme/
[7] https://en.wikipedia.org/wiki/Shredder_1.0

ods, movements, and creators were positioned together – inspired artists to begin creating bricolages from such references as well as extensively quoting them.

Figure 4 – Photomontage by John Heartfield, 1919

What about 'modernism' in the 1910s and 1920s? While the overall emphasis was on originality and novelty, one of the procedures it developed in search of novelty was direct quotations from the vast universe of contemporary visual media that was rapidly expanding at the time. Large headings, for example, and the inclusion of photos and maps made newspapers more visually impactful;

new visually oriented magazines, such as *Vogue* and *Times*, were also launched in 1913 and 1923, respectively; and of course, a new medium of cinema continued to develop.

In response to this visual intensification of mass culture, in the early 1910s Georges Braque and Pablo Picasso began incorporating actual newspaper, poster, wallpaper, and fabric fragments into their paintings. A few years later, John Heartfield, George Grosz, Hannah Hoch, Aleksandr Rodchenko, and a handful of other artists began to develop photo-collage techniques. Photo-collage became another method of creating new media artifacts from existing mass media images.

Contemporary artworks that employ neural networks trained on cultural databases, such as *Unsupervised* or *artificial experiments 1-10*, continue a long tradition of creating new art from *accumulations of images and other media*. In this way, these works of art keep opening up new possibilities for art and its techniques, particularly those of what I referred to as earlier 'database art' (see my article Database as a Symbolic Form, 1998[8]). The introduction of new methods for *reading cultural databases and creating new narratives from them* is part of this expansion.

Thus, *Unsupervised* neither creates collages from existing images, as did modernist artists of the 1920s, nor quotes them extensively, as did postmodern artists of the 1980s. Instead, the group trains a neural network to extract patterns from tens of thousands of MoMA's artworks. The trained net then generates new images that share the same patterns but don't look like any specific paintings. Throughout the course of the animation, we travel through the space of these patterns (e.g., 'latent space'), exploring various regions of the universe of contemporary art. (For a more details about GAN net training methods used by Refik Anadol Studio, see 'Creating Art with Generative Adversarial Network: Refik Anadol's Walt Disney Concert Hall Dreams'[9], 2022).

Pereulkov's Artificial Experiments 1-10 use a different technique to generate new images from an existing image database. He chose only forty paintings by artists who share key characteristics. They developed their oppositional art in late communist society (USSR, 1960s-1980s). They also lived in the same visual culture. In my memories, this society was dominated by two colors: grey (representing the monotony of urban life) and the red of propaganda.

[8] https://manovich.net/index.php/projects/database-as-a-symbolic-form

[9] https://medium.com/@ymingcarina/creating-art-with-generative-adversarial-network-refik-anadols-wdch-dreams-159a6eac762d

In addition, Pereulkov chose paintings that share something else: 'I chose, as a rule, paintings that conceptually relate in some way to the canvas – or to the space on it. I obtained the painting "New Accordion" from Kabakov, which features paper applications on top of the canvas" (my personal communication with Pereulkov, 04/16/2023). Pereulkov also crafted custom text descriptions of each painting used for fine-tuning the Stable Diffusion model. To teach the model the specific visual languages of the chosen artists, he added terms such as 'thick strokes', 'red lighting', 'blue background', and 'flat circles' to these descriptions.

Clearly, each of these steps represents a conceptual and aesthetic decision. In other words, the key to the success of Artificial Experiments 1-10 is the creation of such a database. This work demonstrates how fine-tuning an existing neural network that was trained on billions of image and text pairs (such as Stable Diffusion) can make this network follow artists' ideas; the biases and noise of such a massive network can be overcome and minimized, and do not need to dominate our own imagination.

4. *From Representation to Prediction*

Historically, humans created images of existing or imagined scenes by a number of methods, from manual drawing to 3D CG (see below for explanation of the methods). With AI generative media, a fundamentally new method emerges. Computers use large datasets of existing representations in various media to predict new images (still and animated).

One can certainly propose different historical paths leading to visual generative media today, or divide one historical timeline into different stages. Here is one such possible trajectory:

1. Creating representations manually (e.g. drawing with variety of instruments, carving, etc). More mechanical stages and parts were sometimes carried out by human assistants typically training in their teacher's studio – so there is already some delegation of functions.

2. Creating manually but using assistive devices (e.g. perspective machines, camera lucida). From *hands* to *hands + device*. Now some functions are delegated to mechanical and optical devices.

3. Photography, x-ray, video, volumetric capture, remote sensing, photogrammetry. From *using hands* to *recording information using machines*. From *human assistants* to *machine assistants*.

4. 3D CG. You define a 3d model in a computer and use algorithms that simulate effects of light sources, shadows, fog, transparency, translucency, natural textures, depth of field, motion blur, etc. From r*ecording* to *simulation*.

5. Generative AI. Using media datasets to predict still and moving images. From *simulation* to *prediction.*

'Prediction' is the actual term often used by AI researchers in their publications describing visual generative media methods. So, while this term can be used figuratively and evocatively, this is also what actually happens scientifically when you use image generative tools. When working with a text-to-image AI-model, the neural network attempts to predict the images that correspond best to your text input. I am certainly not suggesting that using all other already accepted terms such as 'generative

media' is inappropriate. But if we want to better understand the difference between AI visual media synthesis methods and other representational methods developed in human history, employing the concept of 'prediction' and thus referring to these AI systems as 'predictive media' captures this difference well.

5. *Media Translations*

There are several methods for creating 'AI media'. One method transforms human media input while retaining the same media type. Text entered by the user, for example, can be summarized, rewritten, expanded, and so on. The output, like the input, is a text. Alternatively, in the image-to-image generation method, one or more input images are used to generate new images.

However, there is another path that is equally intriguing from the historical and theoretical perspectives. 'AI media' can be created by automatically 'translating' content between media types. This is what happens, for example, when you are using Midjoiurney, Stable Diffusion or other AI image generator service and enter a text prompt, and AI generates one or more images in response. Text is 'translated' into an image.

Because this is not a literal one-to-one translation, I put the word 'translation' in quotes. Instead, input from one medium instructs a neural network to predict the appropriate output from another. Such input can also be said to be 'mapped' to some outputs in other media. Text is mapped into new styles of text, images, animation, video, 3D models, and music. The video is converted

into 3D models or animation. Images are 'translated' into text, and so on. Text-to-image method translation is currently more advanced than others, but various forms will catch up eventually.

Translation (or mapping) between one media and another is not a new concept. Such translations were done manually throughout human history, often with artistic intent. Novels have been adapted into plays and films, comic books have been adapted into television series, a fictional or non-fictional text was illustrated with images, etc. Each of these translations was a deliberate cultural act requiring professional skills and knowledge of the appropriate media. Some of these translations can now be performed automatically on a massive scale thanks to artificial neural networks, becoming a new means of communication and culture creation. Of course, artistic adaptation of a novel into a film by a human team and automatic generation of visuals from novel text by a net is not the same thing, but for many more simple cases automatic media translation can work well. What was once a skilled artistic act is now a technological capability available to everyone. We can be sad about everything that might be lost as a result of the automation – and democratization – of this critical cultural operation: skills, something one might call 'deep artistic originality' or 'deep creativity', and so on. However, any such loss may be only temporary if the abilities of 'culture AI' are, for example, even further improved to generate more original content and understand context better.

Because the majority of people in our society can read and write in at least one language, text-to-another media methods are currently the most popular. They include text-to-image, text-to-animation, text-to-3D, and text-to-music models. These AI tools can be used by anyone who can write, or by using readily available translation software to create a prompt in any of the language these tools understand best at a given point. However, other media mappings can be equally interesting for professional creators. Throughout the course of human cultural history, various translations between media types have attracted attention. They include translations between video and music done by VJs in clubs; long literary narratives turned into movies and television series; texts illustrated with images in various media such as engravings; numbers turned into images (digital art); texts describing paintings (ekphrasis tradition, which began in Ancient Greece), mappings between sounds and colors (especially popular in modernist art); etc.

The continued development of AI models for mappings between all types of media, without privileging text, has the po-

tential to be extremely fruitful, and I hope that more tools will be able to accomplish this. Such tools will be very useful both to professional artists and other creators alike. However, being an artist myself, I am not claiming that future 'culture AI' will be able to match, for example, innovative interpretations of Hamlet by avant-garde theatre directors such as Peter Brook or astonishing abstract films by Oscar Fishinger that explored musical and visual correspondences. It is sufficient that new media mapping AI tools stimulate our imagination, provide us with new ideas, and enable us to explore numerous variations of specific designs.

6. *The Stereotypical and the Unique*

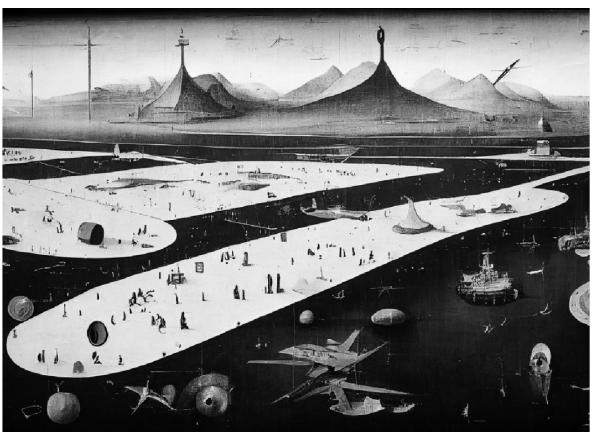
Both the modern human creation process and the predictive AI generative media process seem to function similarly. A neural network is trained using unstructured collections of cultural content, such as billions of images and their descriptions or trillions of web and book pages. The neural net learns associations between these artifacts' constituent parts (such as which words frequently appear next to one another) as well as their common patterns and structures. The trained net then uses these structures, patterns, and 'culture atoms' to create new artifacts when we ask it to. Depending on what we ask for, these AI-created artifacts might closely resemble what already exists or they might not.

Similarly, our life is an ongoing process of both supervised and unsupervised cultural training. We take art and art history courses, view websites, videos, magazines, and exhibition catalogues, visit museums, and travel in order to absorb new cultural information. And when we 'prompt' ourselves to make some new cultural artifacts, our own biological neural and networks (infinitely more complex than any AI nets to date) generate such artifacts based on what we've learned so far: general patterns we've observed, templates for making particular things, and often concrete parts of existing artifacts. In other words, our creations may contain both exact replicas of previously observed artifacts and new things that we represent using templates we have learned, such as golden ratio or use of complementary colors.

AI neural nets used for image generation frequently have a default 'house' style. This is the actual term used by MidJourney developers). If one does not specify a style explicitly, the AI will generate it using this 'default' aesthetic.

Figure 5 – Examples generated in Midjourney version 4 using text prompt "morning sky."

To steer away from this default, you need to add terms to your prompts specifying, a description of the medium, the kind of lighting, the colors and shading, or a phrase like "in the style of" followed by the name of a well-known artist, illustrator, photographer, fashion designer, or architect. Here are two examples of such prompts I made, and the images that Midjourney generated from these prompts. The terms used to define particular style characteristics are highlighted.



Prompt 1:
"giant future 1965 modern airport in Siberia made from water and ice, painted on large wood panel by Hieronymus Bosch, bright pastel colors with white highlights, 23f lens, very detailed --ar 4:3 --s 1250 – test"
(Image generated with Midjourney v3)

Prompt 2:
"Photo of two Russian high-school students, clear skin, very soft studio light, 50mm lens, monochrome, silver tones, high quality, ultra realistic --v 4 --q 2"
(Image generated with Midjourney v4)

This image also illustrates the point I am making later in the chapter: "AI frequently generates new media artifacts that are more stereotypical or idealized than what we intended".

Because it *can simulate many thousands of already-existing aesthetics and styles and interpolate between them to create new hybrid*s, AI is more capable than any single human creator in this regard. However, at present, skilled and highly experienced human creators also have a significant advantage. Both humans and artificial intelligence are capable of imagining and representing nonexistent and existing objects and scenes alike. Yet, unlike AI image generators, human-made images can include very particular content, unique minuscule details, and distinctive aesthetics way that is currently beyond the capabilities of AI. In other words, today a large group of highly skilled and experienced illustrators, photographers, and designers can represent everything a trained neural net can do (although it will take much much longer), but t*hey can also create objects, compositions, or aesthetics that the neural net cannot do at this time. Equally importantly, they can picture unique objects, faces, compositions, and so on – as opposed to often more common-place or idealized versions generated by AI.*

What is the cause of this aesthetic and content gap between human and artificial creators? 'Cultural atoms', structures, and patterns in the training data that occur most frequently are very successfully learned during the process of training an artificial neural network. In the 'mind' of a neural net, they gain more importance. On the other hand, 'atoms' and structures that are rare in the training data or may only appear once are hardly learned or not even parsed at all. They do not enter the artificial culture universe learned by AI. Consequently, when we ask AI to synthesize them, it is unable to do so.

Due to this, text-to-image AIs such as Midjourney, Stable Diffusion or RunwayML are not currently able to generate drawings in my style, expand my drawings by adding newly generated parts, or replace specific portions of my drawings with new content drawn in my style (e.g, they can't perform useful 'outpainting' or 'inpainting' on the digital photos of my drawings.) Instead, these AI tools generate more generic objects than what I frequently draw or they produce something that is merely ambiguous yet uninteresting.

I am certainly not claiming that the style and the world shown in my drawings is completely unique. They are also a result of specific cultural encounters I had, things I observed, and things I noticed. But because they are uncommon (and thus unpredictable), AI finds it difficult to simulate them, at least without additional training using my drawings.

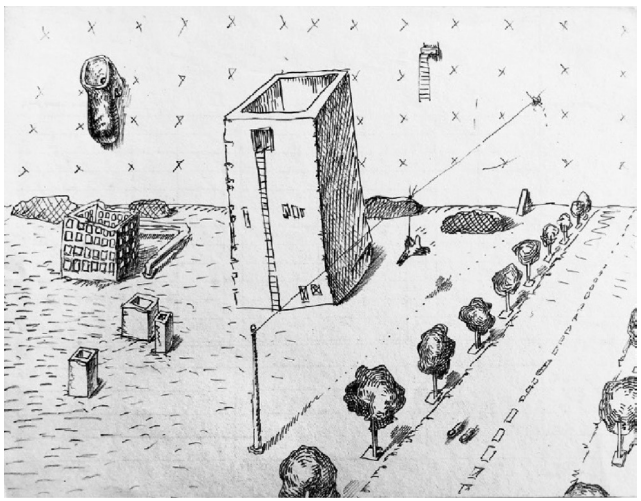


Figure 6 – Lev Manovich, untitled drawing, pen on paper, 1981-1982.

One of my attempts to generate a new version of this image with Stable Diffusion AI, Fall 2022.

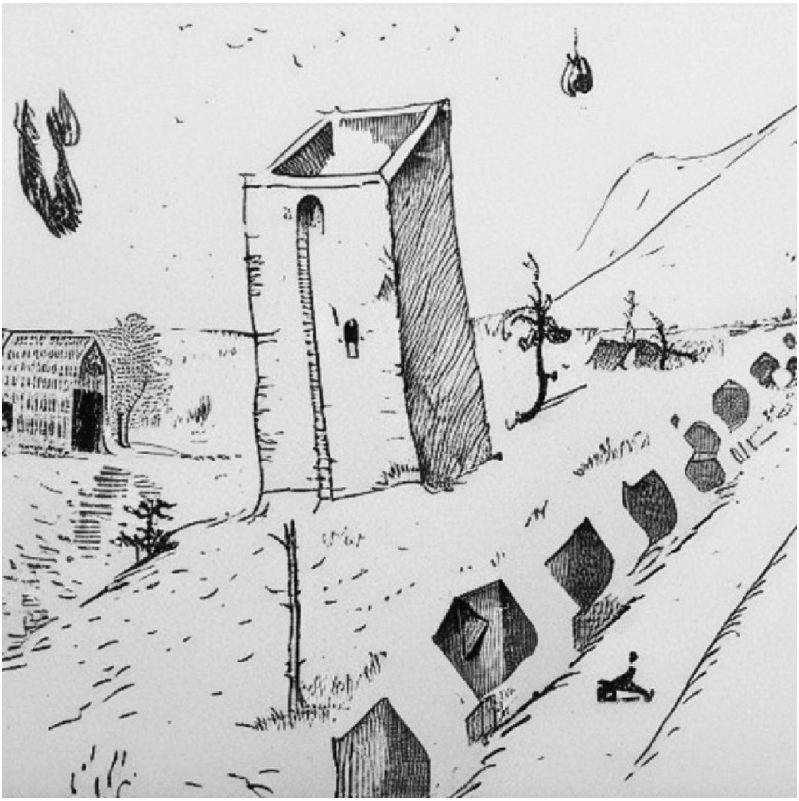


Figure 7 – Stable Difffusion generated figure

Here we encounter what I see as *the greatest obstacle creators face when using AI generative media: AI frequently generates new media artifacts that are more stereotypical or idealized than what we intended*.

This can affect any image dimensions – elements of content, lighting, crosshatching, atmosphere, spatial structure, and details of 3D shapes, among others. Occasionally it is immediately apparent, in which case you can either attempt to correct it or disregard the results. Very often, however, such 'substitu-

tions' are so subtle that we cannot detect them without extensive observation or, in some cases, the use of a computer to quantitatively analyze numerous images. In other words, new AI generative media models, much like the discipline of statistics since its inception in the 18th century and the field of data science since the end of the 2010s, deal well with frequently occurring items and patterns in the data but do not know what to do with the infrequent and uncommon. We can hope that AI researchers will be able to solve this problem in the future, but it seems so fundamental that we should not anticipate a solution immediately.

7. *Subject and Style*

In the arts, the relationship between 'content' and 'form' has been extensively discussed and theorized. This brief section does not attempt to engage in all of these debates or to initiate discussions with all relevant theories. Instead, I would like to consider how these concepts play out in AI's 'generative culture'. However, instead of using content and form, I'll use a different pair of terms that are more common in AI research publications and online conversations between users: *subject* and *style*.

At first glance, AI media tools appear capable of clearly distinguishing between the subject and style of any given representation. In text-to-image models, for instance, you can generate countless images of the same subject. Adding the names of specific artists, media, materials, and art historical periods is all that is required for the same subject to be represented differently to match these references. Photoshop filters began to separate subject and style as soon in the 1990s, but AI generative media tools are more capable. For instance, if you specify 'oil painting' in your prompt, simulated brushstrokes will vary in size and direction across a generated image based on the objects depicted. AI media tools appear to 'understand' the semantics of the representation as opposed to earlier filters that simply applied the same transformation to each image region regardless of its content. For instance, when I used "a painting by Malevich" and "a painting by Bosch" in the same prompt, Midjourney generated an image of space that contained Malevich-like abstract shapes as well as many small human and animal figures like in popular Bosch paintings that were properly scaled for perspective.

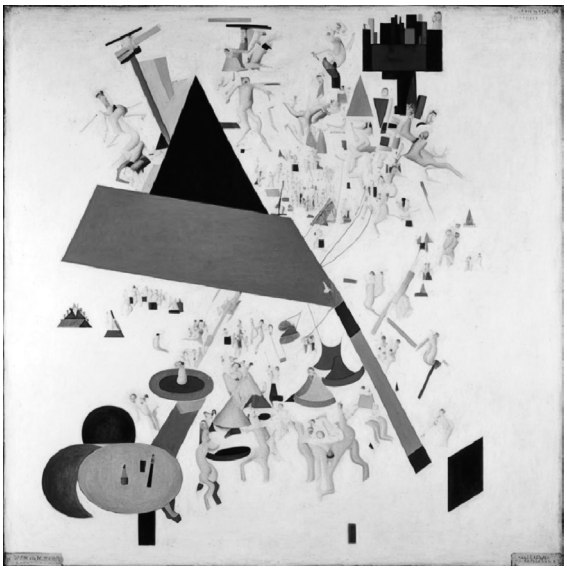


Figure 8 – Image generated in Midjourney using prompt "painting by Malevich and Bosch", Fall 2022.

AI tools routinely add content to an image that I did not specify in my text prompt in addition to representing what I requested. This frequently occurs when the prompt includes 'in the style of' or 'by' followed by the name of a renowned visual artist or photographer. In one experiment, I used the same prompt with the Midjourney AI image tool 148 times, each time adding the name of a different photographer. The subject in the prompt remained mostly the same – an empty landscape with some buildings, a road, and electric poles with wires stretching into the horizon. Sometimes adding a photographer's name had no effect on the elements of a generated image that fit our intuitive concept of style, such as contrast, perspective, and atmosphere. But every now and again, Midjourney also modified the image content. For example, when well-known works by a particular photographer feature human figures in specific poses, the tool would occasionally add such figures to my photographs. (Like Malevich and Bosch, they were transformed to fit the spatial composition of the landscape rather than mechanically duplicated.) Midjourney has also sometimes changed the content of my image to correspond to a historical period when a well-known photographer created his most well-known photographs.

According to my observations, when we ask Midjourney or a similar tool to create an image in the style of a specific artist, and the subject we describe in the prompt is related to the artist's typical subjects, the results can be very successful. However, when the subject of our prompt and the imagery of this artist are very different, 'rendering' the subject in this style frequently fails.

Using prompt "by Caspar David Friedrich --v 5" in Midjourney generates images that capture

the artist's style sufficiently well. Source: https://www.midlibrary.io/styles/caspar-david-friedrich.

Figure 9 – Figure generated with MIdjourney

Using prompt "decaying peonies by Caspar David Friedrich" in Midjourney generates images that simulate important features of artist's style such as combinations of cool colors and dramatic atmosphere. But in other ways, generated images depart significantly from the artist's style. The types of lines, rendering of details, and symmetrical compositions in these AI images would never appear in actual Friedrich's paintings. AI can also often insert some generic looking objects, such as the rock formations in the upper right corner of first image.

Figure 10 – Figures generated with Midjourney

To summarize, in order to successfully simulate a given visual style using current AI tools, you may need to change the content you intended to represent. *Not every subject can be rendered successfully and satisfyingly in any style*. Additionally, AI can often successfully learn some features of artist's style but not others.

These observations, I believe, complicates the binary opposition between the concepts of 'content' and 'style'. For some artists, AI can extract at least some aspects of their style from examples of their work and then apply them to different types of content. But for other artists, it seems, their style and content cannot be separated. (This is only my initial observation to be developed further after doing more experiments in Midjourney.)

For me, these kinds of observations and reflections are one of the most important reasons for using new media technologies like AI generative media and learning how they work. Of course, as a practicing artist and art theorist, I had been thinking about the relationships between subject and style (or content and form) for a long time but being able to conduct systematic experiments like the one I described brings new ideas and allows us to *look back at cultural history and art in new ways*.

References

Manovich, Lev. (1988) Database as a symbolic form. URL = https://manovich.net/index.php/projects/database-as-a-symbolic-form

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., … & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International conference on machine learning* (pp. 8821-8831). Pmlr.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

Vitz, P.C., & Glimcher, A.B. (1985). Modern art and modern science: The parallel analysis of vision. *Journal of Aesthetics and Art Criticism*, *43*(3).

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221513

Sulla natura dell'immagine: Language Model e Visual Language Model

Luca Capone[*], Felice Cimatti[**, ***]

ABSTRACT

L'articolo propone una riflessione sulle capacità di traduzione intersemiotica (tra espressioni grafiche e linguistiche) esibite dai Visual Language Model. Partendo da questo progresso tecnico vengono tratte alcune conclusioni sulla comune natura semiotica di immagine e discorso.

La tesi sostenuta è che le espressioni grafiche e linguistiche condividono una comune e costitutiva indeterminatezza, essendo entrambe determinabili sotto molteplici rispetti (potendo assumere molteplici sensi). L'articolo, smarcandosi da una concezione referenzialista del linguaggio, fa leva sulla teoria semiotica del significato. Secondo questa prospettiva, le immagini (come le espressioni verbali) non rimandano direttamente a cose, ma a significati. Da questo punto di vista, la traducibilità tra immagini e parole non risulta un'anomalia, bensì l'ordinario effetto della loro comune radice semiotica. L'apparente "eccedenza" dell'immagine rispetto all'espressione verbale non segnala un'irriducibile superiorità del visibile, ma il fatto che tanto le immagini quanto i segni linguistici sono sempre aperti alla determinazione.

I Visual Language Model sono un caso esemplare di questa dinamica. In questi strumenti, espressioni grafiche e linguistiche sono organizzate in funzione di un comune sistema semiotico (la lingua). La generazione di immagini a partire da testi (e viceversa la produzione di testo a partire da immagini) mostra che la percezione, nella sua forma umana e artificiale, è già sempre mediata da processi di significazione. Formule come "il visibile non è riducibile al dicibile" non descrivono quindi un fatto empirico, ma si limitano a esprimere una *regola grammaticale*, un pregiudizio della riflessione sul rapporto fra immagini e linguaggio.

KEYWORDS

Semiotics, Language Model, Computer Vision, Transmediality

[*] luca.capone@fileli.unipi.it, CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa

[**] felice.cimatti@unical.it, Dipartimento di Scienze Politiche e Sociali, Università della Calabria

[***] I due autori hanno discusso, progettato e scritto il saggio congiuntamente. Tuttavia, i paragrafi 1 e 6 sono da attribuire a Felice Cimatti, mentre i paragrafi 2, 3, 4, 5 sono da attribuire a Luca Capone.

1. *Dalle parole alle immagini, e viceversa*

Figura 1 – Immagine generata con ChatGPT

Il *prompt* verbale, ossia una sequenza sintatticamente sensata, era molto semplice: "produci una immagine di un gatto con due code che cammina sopra un ramo di una quercia in autunno". Il risultato è questa immagine (figura 1). Dalla scrittura all'immagine. Proviamo ora a rovesciare la richiesta: si tratta adesso di descrivere questa stessa immagine, ossia a trasporre una immagine in parole scritte. Ecco la risposta di ChatGPT: "L'immagine rappresenta una scena autunnale molto suggestiva e stilizzata, quasi pittorica. Al centro della composizione c'è un gatto tigrato dal manto marrone scuro con striature nere. Il gatto è ritratto in movimento, mentre cammina con grazia e sicurezza su un ramo spesso e nodoso di un albero. Ha uno sguardo attento e curioso rivolto verso sinistra, come se avesse percepito qualcosa nei dintorni. Un elemento particolarmente distintivo è la coda del gatto, che si biforca in due punte fluenti, creando una forma elegante e quasi fantastica, dando un tocco surreale o fiabesco all'immagine. Lo sfondo è dominato da foglie di quercia arancioni e giallo-oro, tipiche dell'autunno. Queste foglie riempiono l'ambiente, creando un'atmosfera calda e accogliente, ma anche leggermente magica".

In questo articolo ci poniamo una classica domanda filosofica, una domanda relativa alle condizioni di possibilità di questo passaggio, apparentemente fluido e 'semplice', dalla parola all'immagine e viceversa. Com'è possibile trasformare una descrizione verbale in una immagine? Com'è possibile fare l'operazione inversa? E soprattutto, *perché* è possibile? Che l'immagine del nostro esempio sia stata essa stessa generata da una intelligenza artificiale non cambia la natura del problema, il sistema (ci riferiamo con questo termine neutro al *software* capace di compiere queste operazioni) è in grado

di descrivere qualunque immagine, sia prodotta da esseri umani che artificiali, non c'è differenza. Come spesso accade in filosofia, la questione delle origini non risulta risolutiva, perché o le origini sono sempre operative, e allora non sono propriamente originarie, oppure non ci spiegano perché dovrebbero essere efficaci anche nel presente.

La questione che ci poniamo in questo lavoro è invece rilevante perché spesso si ritiene, o si crede di ritenere, che l'immagine sia sostanzialmente diversa dalle tipiche entità del linguaggio verbale, ad esempio nomi e descrizioni; secondo un modo preteorico di intendere l'immagine questa rientrerebbe in un campo di fenomeni diverso da quello del linguaggio. Il punto in questione non è quello della differenza fra segni arbitrari e segni iconici, perché come sappiamo da Peirce anche questi ultimi sono, a loro modo, arbitrari (la nozione di somiglianza non è una nozione primitiva). Ma questo appunto significa – di questo ci occuperemo in questo lavoro – che fra immagini e parole esiste una analogia semiotica di fondo. In effetti l'esempio con cui abbiamo aperto questo articolo mostra che forse tutta questa differenza non c'è, perché altrimenti come si sarebbe potuta trasformare una immagine in parole, e viceversa? Allo stesso tempo una antica tradizione di pensiero, o forse meglio una tenace sensazione, ci lascia pensare che ci sia comunque 'qualcosa', nella immagine, che la corrispondente descrizione verbale non riesce a catturare. Il visibile non è riducibile, almeno non del tutto, al dicibile. Un'affermazione che per un verso è evidente, per un altro, però, non convince fino in fondo, e un caso come quello da cui hanno preso spunto queste considerazioni sembra confermarlo. Il problema che proviamo a discutere in questo lavoro, allora, è il seguente: "il visibile non è riducibile al dicibile" è una descrizione di che cos'è il visibile, oppure "il visibile non è riducibile al dicibile" è una definizione del visibile?

La posta in gioco è allora la natura dell'immagine. La 'traducibilità' fra immagini e parole va nella direzione di una loro comune natura semiotica. È possibile passare dalle prime alle seconde perché si tratta sempre di segni, sia pure diversi. Ed è inscritta nella natura dei segni che possano trasformarsi, in modo più o meno soddisfacente, gli uni negli altri; da questo punto di vista la prestazione di ChatGPT sarebbe, almeno in linea di principio comprensibile (non ci esprimiamo sulla complessità algoritmica dell'operazione che, nonostante sia sempre evocata, non è pertinente per la nostra discussione: in queste pagine seguiamo la prospettiva del Wittgenstein delle *Ricerche filosofiche*, secondo la quale 'sostanza = funzionamento'). Se l'immagine è un segno allora rimanda, in

un modo o nell'altro, a qualcos'altro, l'oggetto a cui il segno si riferisce, secondo il classico modello semiotico dell'*aliquid stat pro aliquo*. Ma se l'immagine è soprattutto un segno allora la differenza fra segni verbali e non verbali, come appunto il caso dell'immagine, è meno importante di ciò che li accomuna, la comune natura di segni, cioè di qualcosa che rappresenta qualcos'altro. Si apre a questo punto un ulteriore problema, perché ciò a cui il segno si riferisce – il corrispondente oggetto; ad esempio, ciò che intendevamo con il *prompt* "produci una immagine di un gatto con due code che cammina sopra un ramo di una quercia in autunno" – può essere inteso in due modi completamente diversi: secondo la tradizione che possiamo chiamare 'referenzialista', l'oggetto a cui il segno si riferisce è propriamente una cosa, o al più una classe di cose (De Mauro 1970, p. 42). Prendiamo il caso dell'immagine qui sopra: secondo questa concezione il segno-ritratto si riferisce ad un gatto effettivo su un ramo di una quercia. Secondo un'altra concezione – quella propriamente semiotica, e che vale sia per la tradizione che risale a Peirce che per quella che rimanda a Saussure – l'oggetto a cui il segno si riferisce non è una cosa, cioè un'entità non semiotica, bensì un'entità, che qui chiameremo 'significato', *anch'essa semiotica*. Mentre nella tradizione 'referenzialista' fin qui richiamata il significato del segno si colloca al di fuori della semiosi, in quella semiotica che le si contrappone, il significato di ogni evento segnico rimane del tutto all'interno della semiosi.

È evidente che il modo di intendere la portata semiotica delle immagini generate dall'IA cambia completamente a seconda di quale modello di semiosi si adotti. Nel modello 'referenzialista' un'immagine generata da un qualunque sistema di IA non potrà essere intesa come un segno in senso pieno (e quindi nemmeno un'immagine in senso effettivo), appunto perché i modelli che l'hanno 'generata' non 'sanno' nulla del mondo non semiotico (con una terminologia antica, la loro è solo una conoscenza *de dicto*, non *de re*). Secondo questa prospettiva, ad esempio, «We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning […] We take meaning to be the relation between the form and something external to language» (Bender e Koller 2020). Se invece collochiamo un qualunque *Visual Language Model* (VLM) nel quadro teorico di Peirce o di Saussure (trascurando le ovvie differenze fra questi due approcci), non avremo nessuna difficoltà a riconoscerle come immagini a pieno titolo, perché in questa tradizione di pensiero un'immagine 'funziona' come immagine, e quindi come dispositivo semiotico, a prescindere da qualunque riferimento a presunte entità esterne non semiotiche.

Prendiamo, ad esempio, il caso di Peirce: nel suo modello semiotico l'immagine rientra nella classe delle icone, e l'icona è un segno esattamente allo stesso titolo di un simbolo o un indice, e per tutti questi segni vale che il loro 'significato' è costituito dal cosiddetto 'interpretante' che lo collega al suo 'oggetto': il carattere specifico dell'interpretante, una caratteristica che non smette di stupire la tradizione 'referenzialista', è di essere a sua volta un segno. È vero che per lo stesso Peirce l'icona svolge un ruolo particolare nel processo conoscitivo, quello iniziale, abduttivo, collocandosi quindi ai bordi estremi della semiosi (non a caso l'abduzione si collega alla categoria della Primità). Tuttavia un'icona è un segno come gli altri. Una delle conseguenze più indigeribili – per la prospettiva referenzialista – di questa posizione è che rende inutilizzabili o più semplicemente superflue le nozioni di 'verità' e 'falsità' per rendere conto delle immagini. L'immagine del gatto con le due code è 'falsa' nel mondo dei gatti di cui abbiamo esperienza (anche se in rete si trovano non poche 'immagini' di gatti 'veri' con due code), ma non lo è in 'altri' mondi, come quello letterario ad esempio. Ma è proprio questo il punto, l'immagine è davvero, e principalmente, un segno? Si può in effetti pensare in modo diverso l'immagine, senza includerla in tutto e per tutto nella classe più generale dei segni. O meglio, si può continuare a considerarla in qualche modo un segno – perché nell'immagine ci viene 'naturale' vedere una rappresentazione di qualcos'altro, e senza questa caratteristica non potremmo parlare nemmeno di immagini – senza però identificarla completamente con un segno. In questo modo se ne mantiene la natura rappresentativa, senza però assimilarla in tutto e per tutto ai segni più 'classici', come gli indici e i simboli. Secondo questa prospettiva l'immagine è rappresentativa, altrimenti non sarebbe una immagine, senza però essere del tutto assimilabile ad un segno. Questa distinzione, ammesso che sia praticabile, dovrebbe servire a garantire all'immagine uno statuto particolare, possibilmente autonomo, rispetto ai segni linguistici. L'immagine avrebbe allora uno statuto affatto peculiare, allo stesso tempo segnico, ma anche non segnico.

La posta in gioco, in effetti, è il rapporto fra l'immagine e il linguaggio verbale. L'immagine è autonoma, in tutto o in parte, rispetto ai segni linguistici? Da come si risponde a questa domanda derivano atteggiamenti molto diversi rispetto alla generazione di immagini da parte di VLM. Per chi ritiene che l'immagine non rientri nel campo dei segni, allora le immagini generate dall'IA sono perfetti esempi di immagine, proprio perché il sistema che le genera non 'sa' nulla del mondo cosiddetto 'reale'. Ma la 'facilità' con cui una intelligenza artificiale come ChatGPT passa dalle parole

all'immagine e viceversa, rende poco plausibile questa posizione. Al contrario, è proprio questa facilità che rende molto convincente l'assimilazione delle immagini al campo della semiosi. Qual è la prestazione specifica della semiotica, in fondo, se non proprio di porre in questione la distinzione fra *de dicto* e *de re*?

Se l'immagine è un segno in senso pieno, come ritiene Peirce, allora nel passaggio dall'una all'altro non si dovrebbe perdere, almeno in linea di principio, alcuna informazione, e quindi avremo una sostanziale equivalenza fra una immagine e la sua corrispondente descrizione verbale: l'immagine sarebbe una sorta di discorso senza parole, e una descrizione sarebbe una sequenza di parole senza figure. Questa formulazione è evidentemente paradossale, tuttavia al fondo i programmi che trasformano descrizioni verbali più o meno accurate in corrispondenti immagini e, al contrario, trasformano immagini in descrizioni funzionano in base a questo fondamentale principio semiotico: le immagini sono segni, e siccome il modello della semiosi è quello simbolico (cioè quello dei segni linguistici, come sosteneva Barthes), ne deriva la sorprendente conclusione che la trasposizione di immagini in parole e parole in immagini, per quanto possano sembrare entità appartenenti a generi molto diversi fra loro, non pone particolari problemi. In questo senso la 'novità' dei sistemi di IA non è poi così innovativa, dal momento che in fondo non fanno che rendere esplicito il comune fondamento semiotico di immagini e parole. Come scrive Garroni, non rimane che accettare che "*percezione*" – in quanto attività che produce immagini – "*e linguaggio si condizionino a vicenda*, che l'uno supponga l'altra e quella supponga questo, che essi insomma costituiscano una stretta correlazione" (Garroni 2005, p. 41).

Tuttavia questa apparente semplicità si basa su un'operazione, che è allo stesso tempo economica linguistica e semiotica, l'annotazione di dati. Nel caso più brutale si tratta del lavoro più o meno nascosto di migliaia di esseri umani a cui vengono sottoposte delle immagini che devono trasporre in parole: l'operazione nota nella nostra tradizione culturale come *ekphrasis* (*εκφρασις*). Si prenda il caso classico del Fedro, quando Socrate descrive il paesaggio sotto il platano dove vuole ascoltare le parole di Fedro: "Per Era, che bel posto! Quel platano così alto e pieno di fronde; e quell'agnocasto, lui pure così alto, con quell'ombra meravigliosa; è in piena fioritura e riempie di profumi tutto intorno. E questa sorgente davvero graziosa che scorre sotto il platano! Le sue acque sono proprio fresche, ne fanno prova i miei piedi. […]. Senti poi, te ne prego, la brezza di questo luogo quanto è amabile e davvero piacevole: è il canto estivo che risponde al coro delle cicale" (*Phaedrus* 230 c-d).

Nei termini attuali, anche se potrà sembrare irrispettoso, Platone nello scrivere il *Fedro* – in particolare quello che sta vedendo Socrate sotto il platano – sta appunto compiendo un'operazione di 'annotazione': sta trasponendo un'immagine (mentale) di un luogo in un discorso, ossia una sequenza di parole, di *note* appunto, che trasformano una scena visiva in una sequenza di parole. Messa in questi termini sembrerebbe immediatamente evidente la sproporzione fra la ricchezza della scena vista e partecipata con tutto il corpo e la sua resa verbale, che per quanto accurata sembra appunto essere affatto sottodeterminata rispetto a quella profusione di colori, profumi ed esperienze tattili.

Ma è proprio così? È vero, come sostiene una formula mille volte ripetuta, che "a picture is worth a thousand words"? (Hasenmueller 1989). Questa formula è una descrizione contrastiva dell'immagine rispetto al linguaggio verbale, o una definizione dell'immagine (Cometa 2011)? Secondo Garroni nell'immagine troviamo "*l'unione del determinato e dell'indeterminato*" (Garroni 2005, p. 26), perché per un verso un'immagine è sempre "incompleta" (*Ivi*, p. 21), peraltro però – e proprio grazie all'interazione con il linguaggio verbale – è anche sempre ulteriormente determinabile, precisabile. In effetti, per tornare ai programmi che 'generano' immagini, più il *prompt* è dettagliato più l'immagine 'prodotta' sarà corrispondentemente determinata. Tuttavia, e questo è un altro punto di vicinanza fra immagine e parola, "determinato e indeterminato sono complementari. Là dove c'è una qualche determinazione deve esserci anche una correlativa indeterminazione" (*Ivi*, p. 33). Ci troviamo così di nuovo in quello che Garroni chiama efficacemente "l'enigma della percezione" – in quanto facoltà delle immagini: "l'immagine è per un verso determinata e per un altro indeterminata" (*Ivi*, p. 25). Quello che l'immagine con cui abbiamo aperto queste considerazioni mostra allora è proprio questa situazione: vicinanza sostanziale con la parola – quella vicinanza che permette la 'traducibilità' dell'una all'altra e viceversa – ma anche differenza, nel senso che l'immagine è comunque incompleta, costitutivamente incompleta, sempre ulteriormente determinabile. Ma questa stessa differenza si rivela in realtà una nuova vicinanza, perché anche la parola è altrettanto indeterminata. Se ci liberiamo di quella che Eco chiamava "la fallacia referenziale" (Eco 1975, §2.5.1) – cioè dalla ricerca di una cosa come terminale della semiosi – ci accorgiamo così che le immagini generate dall'IA condividono con i testi prodotti dalla IA uno stesso doppio carattere: si tratta di costrutti allo stesso tempo indeterminati e determinati (o meglio, sempre ulteriormente determinabili).

2. *Il concetto di indeterminatezza*

Nel paragrafo precedente lo statuto delle immagini è risultato conteso fra due prospettive antagoniste. Quella referenzialista, per cui l'immagine trova il suo termine correlativo in un oggetto reale, e quella semiotica, per cui le immagini sono assimilabili a segni che, in quanto tali, rimandano ad altri segni interpretanti. Le prestazioni dei modelli di IA sembrano rendere plausibile la *vicinanza sostanziale* fra parola e immagine, condizione della traducibilità (automatica e non) fra le due. Questa vicinanza lascia aperte due questioni: la prima riguarda le condizioni di possibilità della *traduzione* fra immagini e parole (§3), la seconda è relativa all'eventuale scarto di questo processo di conversione, posto dall'apparente ricchezza dell'immagine, a confronto con la sinteticità del segno linguistico (§4). La risoluzione di entrambi i problemi dipende dalla chiarificazione del carattere *indeterminato* di parole e immagini. Prima di passare alla loro risoluzione è necessario quindi specificare in che modo parole e immagini sono fondamentalmente indeterminate.

2.1 *Indeterminatezza delle immagini*

Il caso più semplice è quello dell'immagine[1]. L'immagine presenta un numero, se non infinito, almeno difficilmente numerabile di tratti potenzialmente pertinenti (cfr. De Mauro 2024, pp. 14-15). Questo numero elevato di tratti costituisce un problema, ovvero come selezionare alcuni di essi ai fini dell'identificazione (o della descrizione) dell'immagine. Il fatto che i tratti siano in numero elevato fa sì che l'immagine possa essere identificata (pertinentizzata) seguendo diverse strategie (prospettive). In altri termini, la medesima immagine può costituire il *senso* di un numero potenzialmente infinito di espressioni: *un gatto comune, un animale in equilibrio su un ramo, allegoria dell'autunno, esempio di gatto irrealistico, immagine dai colori caldi* ecc. Da questo punto di vista l'immagine costituisce una materia aperta, disponibile a molteplici determinazioni, più o meno avventurose. Chiaramente, la materialità dell'immagine rende implausibili alcune pertinentizzazioni, tuttavia, fra quelle ac-

[1] Gli autori sono consapevoli della differenziazione figura-immagine operata da Garroni: l'immagine (interna e dinamica), la figura (esterna e statica). Dato che i modelli di IA hanno a che fare sempre e solo con figure determinate (non con immagini interne), l'articolo si concentra su queste. Tuttavia, nella letteratura la differenziazione garroniana non trova particolare spazio. È stato quindi ritenuto opportuno usare la parola immagine per riferirsi a queste espressioni grafiche determinate. Di conseguenza, l'obiettivo dei seguenti paragrafi è la descrizione del rapporto fra i segni linguistici e queste espressioni grafiche determinate (da qui in avanti *immagini*).

cettabili, nessuna in quanto tale è in grado di imporsi sulle altre. A causa di questa plurivocità di fondo, le immagini sono caratterizzate da Garroni come indeterminate (2005, p. 47, plurinterpretabilità delle figure). È importante sottolineare che nessuna di queste determinazioni possibili è data in quanto tale, ma tutte dipendono da un secondo fattore, ovvero dalla sussunzione dell'immagine da parte di un significato.

2.2 *Indeterminatezza dei segni linguistici*

Per quanto riguarda la parola, la situazione è più delicata a causa della diffusione di quella che è stata definita *fallacia referenziale* (Eco 1975, §2.5.1), ovvero l'idea secondo cui il significato di una parola è l'oggetto per cui essa sta. Da questo punto di vista il significato di una parola come *mela* sarebbe l'oggetto a cui si riferisce. Vale la pena ribadire che una mela determinata può costituire il senso puntuale di una espressione pronunciata *in un determinato contesto* ("passami la mela che sta sul tavolo"), tuttavia la possibilità di riferirsi a oggetti specifici (assumere come sensi oggetti puntuali) dipende da un fenomeno più generale e indeterminato, ovvero dal significato.

Il significato non è un oggetto determinato ma è un concetto, ovvero un termine generale e astratto. Un termine o un'espressione in isolamento (ad esempio "la mela è rossa") non ha un senso determinato, non si riferisce ad alcun oggetto specifico, ma possiede solo un significato (Eco 2016, p. 351). Le espressioni, *in determinati contesti*, possono essere utilizzate per veicolare sensi determinati, riferendosi ad esempio a oggetti specifici, ma sempre sulla scorta di un significato precedentemente dato. La confusione fra significato (astratto e generale) e oggetto o referente (determinato e individuale) richiama quella evidenziata nelle *Ricerche filosofiche* fra significato del nome e portatore (Wittgenstein 2014, §40). È inoltre importante sottolineare che un significato generale non deve necessariamente incarnarsi in sensi referenziali (riferimento a oggetti), ma molti altri usi delle parole sono possibili: raccontare una storia, fornire una definizione, dare un comando ecc.

Eco descrive il significato come un'unità concettuale che una cultura ha differenziato da altre (Eco 1975, §2.6). Questa definizione è importante in quanto consente di concepire il significato non come un'unità in isolamento, ma come parte di un sistema di differenziazioni solidali, le quali articolano il piano del contenuto (o dei significati) di una lingua. Inoltre, queste articolazioni non sono imposte dall'esperienza (così come l'immagine non poteva imporre

la sua identificazione), ma sono ricevute dai parlanti nella lingua che apprendono (cfr. Violi 1997). È opportuno insistere su questi due punti: solidarietà e arbitrarietà delle articolazioni del piano del contenuto. Ad esempio, le tassonomie ortofrutticole, al pari di ogni altra organizzazione del continuum dell'esperienza, sono il frutto dell'opera articolatoria delle lingue storico naturali. Nell'italiano, da un punto di vista rigoroso, la categoria della frutta seleziona dei tratti in funzione di criteri di classificazione botanici, per cui è *frutta* il prodotto della modificazione dell'ovario della pianta. Questa classificazione assimila le mele alle melanzane ed esclude le fragole e le angurie. In questo caso, l'accezione botanica di *frutta* ha selezionato un criterio, fra i molti possibili, per articolare una parte del piano del contenuto. Tuttavia, nell'uso comune il significato *frutta* si attesta su criteri culinari, assimilando mele, fragole e angurie, ma escludendo la melanzana. In breve, l'accezione comune pertinentizza altri tratti dell'esperienza, la quale si offre a molteplici determinazioni possibili (Eco 2016, p. 70). Le due accezioni di *frutta* presentate esemplificano il modo in cui i significati articolano il piano del contenuto in maniera *arbitraria*, potendo cioè scegliere fra diverse possibili strategie. Inoltre, l'esempio evidenzia la solidarietà fra le articolazioni. Ovvero, cambiando il criterio di pertinetizzazione del contenuto di un significato (*frutta*), cambiano i rapporti reciproci fra gli altri significati del sistema (*mela, anguria, melanzana* ecc.). Riassumendo, il piano del contenuto della lingua delinea unità culturali, distinguendo i contenuti che una cultura ha eletto come elementi distinti rispetto ad altri (Eco 1975, §2.6). Queste unità sono solidali, nel senso che si determinano in funzione dei rapporti reciproci. Al tempo stesso sono arbitrarie, nel senso che altre strategie di articolazione del piano del contenuto sarebbero state possibili. Lo stesso discorso potrebbe esser fatto per il significato *gatto*, il quale isola all'interno della famiglia dei felidi una specifica sottocategoria, in base a criteri eterogenei selezionati arbitrariamente.

Questi esempi, oltre a chiarire lo statuto infralinguistico (o semiotico) del significato, ne evidenziano anche il carattere generale e indeterminato. In primo luogo, un significato può pertinentizzare *alcuni* tratti, lasciandone molti altri indeterminati. Il significato *gatto* annoverà molti tratti utili all'identificazione di questi animali (un certo aspetto, un modo di incedere, nozioni sulle loro abitudini ecc.), lasciandone tuttavia molti altri vaghi (colore) o del tutto non specificati (numero di miagolii medi giornalieri). Questo vale sia per i significati comuni, che per quelli più determinati e specifici, come la definizione geometrica del quadrato, la quale non contiene

specifiche relative alle dimensioni, non pertinenti per la definizione delle forme geometriche e delle loro proprietà. Questo spazio di indeterminazione consente a un significato di incarnarsi in molti sensi determinati (ad esempio riferendosi a molti gatti o quadrati diversi). Infine un significato, sebbene costituisca un termine generale, non necessariamente deve possedere le caratteristiche di una classe, i cui membri devono disporre di un numero di caratteristiche necessarie e sufficienti (De Mauro 2024, p. 104). I significati sono affini al concetto di *famiglia* delle *Ricerche filosofiche* (Wittgenstein 2014, §67), i cui membri sono tenuti insieme da somiglianze eterogenee e intermittenti. Ovvero, i tratti pertinenti di un significato possono giocare ruoli diversi a seconda del contesto, per la produzione di sensi nuovi e diversi (De Mauro 2024, p. 118).

Ora che è stato chiarito il tipo di indeterminatezza che pertiene a significato e immagine è possibile descrivere in che modo queste due entità possono entrare in rapporto reciproco.

3. *Cose, parole e immagini*

In *Immagine, linguaggio, figura*, Garroni (2005) cerca di mettere a fuoco la questione del rapporto fra percezione e linguaggio. Introducendo l'argomento, descrive l'esperienza ordinaria di qualcuno che scorra in rassegna con lo sguardo alcuni oggetti, senza un interesse determinato. Garroni riconosce una certa identità agli oggetti così esperiti, tipica non ancora del concetto (il quale pone l'identificazione sotto un rispetto particolare), ma dell'aggregato. In questo senso la percezione, differenziandosi dal linguaggio e senza rifarsi esplicitamente ad esso, in qualche modo lo anticipa e lo richiede (Garroni 2005, p. 10). La questione del riconoscimento richiama il linguaggio, nel senso che la percezione fornisce il materiale indeterminato, ma determinabile, su cui agisce il linguaggio. "Offre al linguaggio il materiale e la garanzia oggettiva dei significati, ma non ancora significati" (*Ivi*, p. 52). In altre parole, nell'assoluta singolarità e plurideterminabilità dell'esperienza sensibile, la lingua determina un ordine, selezionando certi tratti fra quelli disponibili e trascurandone altri. In questo modo il percepito individuale è ricondotto a concetti generali (Garroni parla di classi o famiglie. *Ivi,* p. 10) consentendo, fra le altre cose, il riconoscimento di gatti e alberi in quanto esemplari dei significati "gatto" e "albero" (cioè sempre dentro un'organizzazione semiotica dell'esperienza). Garroni ovviamente non intende dire che nella percezione non si verifica alcuna individuazione o organizzazione

dell'esperienza. Tuttavia, questa organizzazione percettiva è solo l'inizio, ancora indeterminato, del processo di riconoscimento. L'autore insiste su questo punto approfondendo la differenza fra percezione e linguaggio. Secondo Garroni, la prima è destinata a rimanere legata agli oggetti che di volta in volta le si oppongono e ai loro aspetti sensibili attualmente presenti. D'altro canto, il linguaggio (e specifica, qualsiasi linguaggio, anche quello *iniziale*) attiva necessariamente la "determinazione semantica e intellettuale di quegli oggetti, dei loro aspetti e delle relazioni tra di essi e tra soggetti e oggetti" (*Ivi*, p. 52). Ovvero, il linguaggio riconduce la percezione a un significato generale, il quale è in relazione con tutto il sistema di concetti che la lingua mette a disposizione. In questo modo entra in gioco un sistema di relazioni che allontana l'esperienza dalla sua iniziale prossimità con il sentito e il percepito. In questo caso Garroni non pensa solo alle determinazioni specialistiche, che ad esempio concettualizzano un gatto come un *mammifero carnivoro della famiglia dei felidi*, ma a qualsiasi determinazione che la lingua sia in grado di fornire (anche "mitologica o magica", *Ibid.*). In breve, affinché sia possibile riconoscere un gatto come un *gatto* (e non semplicemente come un aggregato di tratti individuali, al più riconducibili a esperienze passate), è necessaria un'astrazione dall'occorrenza sensibile puntuale, per rapportarla a un significato (termine generale) e metterla in collegamento con il resto dell'orizzonte di senso offerto da una lingua storico naturale. Questo è un lavoro dai risultati non scontati. L'esperienza non può causare autonomamente il riconoscimento, ma è di per sé aperta a possibili determinazioni. Per apprezzare la necessità di questa sussunzione concettuale, basti pensare alle peripezie descrittive di chi si è trovato di fronte a un *qualcosa* che violava il sistema di significati a disposizione fino a quel momento. Sono casi noti quelli dell'ornitorinco e della descrizione del rinoceronte di Marco Polo, riferiti da Eco (2016, p. 77). Sul versante grafico, Gombrich (2002) riporta molti esempi di illustrazioni influenzate da concetti familiari. Castel sant'Angelo rappresentato con i tratti di una Burg da un illustratore tedesco (*Ivi,* p. 79) e diversi altri esempi di animali esotici familiarizzati (*Ivi,* pp. 87-89). In tutte queste storie, i protagonisti hanno ricondotto la loro esperienza puntuale al sistema di significati a loro disposizione.

Da un punto di vista ontogenetico, la forma di organizzazione aggregativa, propria della percezione pre-linguistica, convive a lungo con quella linguistica, esibendo formazioni di compromesso (il linguaggio *iniziale* di cui parla Garroni 2005, p. 52). Ne è un esempio classico il peculiare utilizzo delle parole da parte di bambini molto

piccoli. Questi impiegano un medesimo termine per indicare cose diverse, legate da rapporti eterogenei e talvolta difficili da individuare (Vygotskij 1990, p. 171, riporta la parola 'vau-vau', usata per indicare *un cane*, *un boa di pelliccia*, *una figurina di porcellana raffigurante una ragazza*, *i bottoni di una giacca* ecc. Garroni 2005, p. 54, parla del termine 'scarpa' usato per *mettersi le scarpe*, *alzarsi dal letto*, *andare a passeggio*). Ci vorrà del tempo prima che i bambini imparino ad articolare l'esperienza al modo degli adulti (ovvero secondo concetti).

In base a quanto sostenuto da Garroni, per un essere umano la situazione con l'immagine del gatto con due code (figura 1) sembrerebbe la seguente. L'immagine presenta molteplici tratti selezionabili. La sussunzione linguistica seleziona certi tratti, riportando il percepito a un significato generale e consentendo così il riconoscimento e la descrizione del contenuto. In accordo con la concezione semiotica dell'immagine, questa rimanda non a un oggetto, ma a un significato. Effettivamente l'immagine non corrisponde ad alcun gatto reale, tuttavia la situazione non sarebbe cambiata con una fotografia di un oggetto realmente esistente, come ad esempio del Colosseo. L'immagine esibisce sempre *alcuni* dei tratti pertinenti di un oggetto, che possono essere *riconosciuti e sintetizzati* facendo entrare in gioco un significato. Il fatto che l'immagine rimandi a un significato e non a un oggetto, non esclude ovviamente che un'immagine possa essere usata come espressione per riferirsi a un oggetto concreto, in determinati contesti. Tuttavia, come per le parole, il riferimento ha per condizione un significato generale che deve mediare il riferimento al senso individuale. Fuori da questa sussunzione linguistica l'immagine non è nulla, nelle parole di Garroni: "riconoscere qualcosa o equivale a essere già pronti a parlarne, in forza di qualche linguaggio correlato alla percezione, oppure è un atto senza esito […] destinato ad autoconsumarsi […] non incidente sulla nostra presa di coscienza del mondo, anche solo percettiva" (*Ivi,* p. 59). Per concludere, sembrerebbe che sebbene si possa avere un linguaggio senza immagini, non sarebbe possibile avere immagini senza un linguaggio, a causa della loro fondamentale indeterminazione.

Garroni non si esprime sul caso inverso (dalla parola all'immagine), tuttavia gli esempi già citati di Gombrich possono essere di aiuto. In questo caso la produzione grafica segue degli stilemi codificati nella pratica rappresentativa, i quali possono presentare elementi di sorprendente dipendenza dalla cultura di chi si presta a rappresentare qualcosa. Nell'immagine prodotta dal VLM non è un caso che il gatto rappresentato fosse un Gatto Europeo. La

familiarità con una specifica determinazione del significato *gatto* ("produci una immagine di un gatto con due code che cammina sopra un ramo di una quercia in autunno") ha probabilmente influenzato la produzione grafica del modello.

4. *L'eccedenza dell'immagine*

Nei precedenti paragrafi è stato descritto il modo in cui immagini e linguaggio possono entrare in reciproca relazione. Il rapporto è apparso sbilanciato dal lato del linguaggio, i cui significati sono la condizione del riconoscimento di immagini. Tuttavia una questione rimane in sospeso, ovvero quella dell'eccedenza dell'immagine rispetto alla sua traduzione linguistica. Questa eccedenza non si differenzia da quella che è stata chiamata *indeterminatezza dell'immagine*, ovvero dal fatto che un'immagine può costituire il senso di innumerevoli espressioni. Tuttavia la critica in questo caso sembra malposta, principalmente per due motivi.

Il primo motivo è che il linguaggio non è fondamentalmente incapace di tradure adeguatamente un'immagine, semplicemente non è in grado di rappresentare simultaneamente tutte le innumerevoli possibili prospettive da cui un'immagine può essere *tradotta* linguisticamente. Tuttavia, nemmeno un'immagine è in grado di farlo, l'immagine si presenta sempre sotto un certo rispetto, è *letta* sempre da una qualche prospettiva. La quale può essere mutata, invertita, falsificata, ma sempre in un susseguirsi di interpretazioni che la fissano. Da questo punto di vista, l'immagine costituisce una materia indeterminata ma determinabile, disponibile all'azione formatrice della lingua. Le determinazioni della materia sensibile, in questo caso visiva, possono seguire tratti altamente codificati, come il pelo, le orecchie e le zampe del gatto. Oppure possono optare per strategie non codificate, assumendo tratti della materia come possibili "unità pertinenti di un codice in formazione" (Eco 1975, §2.14.4), come accade per esempio nel caso dell'interpretazione della musica (Capone 2023). Quale che sia la strategia di determinazione adottata, l'immagine di per sé, non sembra possedere alcuna nascosta ricchezza inaccessibile al linguaggio.

Il secondo motivo per cui questa critica non coglie il segno è che potrebbe essere mossa anche nei confronti del medium grafico. Una frase apparentemente semplice come "un gatto con due code che cammina sopra un ramo di una quercia in autunno" presenta un grado di indeterminazione non inferiore a quello dell'immagine. L'espressione potrebbe costituire il senso di un numero indefinito di imma-

gini, che la rappresentino da prospettive e secondo strategie sempre nuove, determinando graficamente il medesimo significato. Non per questo tuttavia sembra ragionevole affermare che la parola presenta una ricchezza inattingibile all'immagine e intraducibile in figure.

A questo punto, anche l'obiezione relativa a un'ipotetica irriducibilità delle immagini al linguaggio sembra risolta. Dai precedenti paragrafi emerge lo statuto segnico delle immagini. Il correlato naturale della percezione umana sembra essere una competenza semiotica sofisticata, come condizione del rapporto che gli esseri umani intrattengo con le immagini e più in generale con la propria percezione. Tuttavia, se la spiegazione di Garroni può essere soddisfacente per gli esseri umani, non sembra del tutto esaustiva per i modelli, i quali non dispongono di una percezione flessibile e di una capacità di organizzazione dell'esperienza percettiva pre-linguistica. Nel prossimo paragrafo verrà analizzato il caso dei VLM.

5. *Visual Language Models*

Garroni postula una percezione flessibile e sofisticata che lavora in stretta correlazione con il linguaggio. Per i Language Model (LM) e i VLM questa possibilità è da escludere. Se la percezione consente di far emergere delle *linee di resistenza* (Eco 2016, p.69) nella materia sensibile (conservandone comunque l'indeterminata forma aggregativa), per un modello la situazione è ancora più incerta. Per un VLM (un modello in grado di ricevere e generare sia immagini che testo) un'immagine si presenta come una matrice numerica, i cui valori determinano l'intensità dei pixel.[2] Questa matrice è costituita da delle serie di numeri ordinati che non offrono nessuna preferenza, nessuna linea di resistenza in vista di un'organizzazione in gruppi di tratti correlabili a significati. In questo magma isotropico deve subentrare un principio di classificazione esterno, affinché questi rapporti numerici possano essere correlati con dei termini significativi. Questo è ciò che succede nei VLM.

Un VLM è fondamentalmente una rete neurale, la quale trasforma il proprio input (testo o immagini) in una serie di vettori.[3]

[2] In questo caso viene dato per scontato che l'immagine sia in bianco e nero, rappresentata quindi da una matrice bidimensionale n x m. I casi di immagini a colori chiamano in causa strutture a più dimensioni, come i *tensori*. Questo non incide sul principio di funzionamento del modello.

[3] I vettori sono la struttura dati attraverso cui le reti neurali rappresentano i propri input. La forza rappresentativa dei vettori sta nel fatto che non rappresentano in isolamento i propri oggetti ma li pongono in uno spazio rappresentativo in funzione di reciproci rapporti con altri elementi (Capone 2021, Capone e Cimatti 2025).

Durante l'addestramento un VLM *apprende* a rappresentare immagini e testi fra loro correlati tramite vettori simili. Questo obiettivo viene raggiunto tramite un addestramento congiunto con immagini correlate a sequenze di parole (esistono diverse tecniche, Ramesh *et al.* 2022, Sharia *et al.* 2022, OpenAI 2023). In questo modo, il modello apprende a relazionare rappresentazioni di elementi visivi e linguistici, sviluppando rappresentazioni congiunte. Un modello di questo tipo può ricevere in input un'immagine e, dopo averla rappresentata vettorialmente, generarne la descrizione linguistica. In una fase successiva, queste rappresentazioni vettoriali vengono utilizzate come condizionamento per l'addestramento di una seconda parte del modello, dedicata alla generazione di immagini (Nichol *et al.* 2022).

Un VLM non costruisce direttamente un'immagine dal nulla, ma procede per sottrazione, eliminando progressivamente del rumore da una matrice inizialmente composta da puro rumore[4]. Il cuore di questo processo è una rete neurale detta *denoiser*, la quale viene addestrata a rimuovere il rumore da immagini solo parzialmente corrotte, ricostruendole.

Al termine dell'addestramento, il modello è in grado di generare un'immagine coerente a partire da un prompt linguistico (vettorialmente rappresentato), rimuovendo iterativamente il rumore da una matrice iniziale completamente casuale. In questo modo il modello procede per sottrazione, facendo emergere nell'immagine differenziazioni significative in funzione di un condizionamento linguistico. Il VLM, non avendo a disposizione una sensibilità vera e propria, appalta tutto il lavoro esplorativo, profilativo, articolatorio della materia sensibile al versante linguistico, al principio di classificazione rappresentato dalla lingua. In questo modo è in grado di realizzare un'immagine partendo da una matrice di rumore indifferenziato.

6. *Conclusioni*

I paragrafi precedenti hanno cercato di dare una risposta a due questioni emerse dall'analisi del rapporto fra immagini e linguaggio nei sistemi di IA: la questione delle condizioni di possibilità che consentono la traduzione di immagini in linguaggio e viceversa; il problema dell'eccedenza dell'immagine rispetto alla sua resa linguistica. La risoluzione di queste due questioni consente ora di chiarifi-

[4] In breve, il modello riceve (oltre alle istruzioni testuali)un'immagine sgranata, composta da pixel casuali, in cui non è possibile identificare alcun contenuto (in termini semiotici, una materia amorfa).

care il senso di alcune proposizioni introdotte nel primo paragrafo, le quali non consentivano di porre il problema del rapporto fra immagini e linguaggio in termini produttivi.

Proviamo a trasformare la formula "a picture is worth a thousand words" in una domanda: "Is a picture worth a thousand words?". Un'immagine può riassumere il senso di numerosi enunciati, ma allo stesso tempo un enunciato può sintetizzare il senso complessivo di innumerevoli immagini. Il punto da ribadire è che il significato di una espressione non è meno indeterminato del significato di una immagine e, allo stesso modo, può essere determinato quanto può esserlo una immagine. La traducibilità di una immagine in una descrizione verbale, e viceversa, rimanda allora a questa caratteristica fondamentale dei segni: il loro significato è sempre determinato/indeterminato, ossia sempre indefinitamente determinabile.

Riguardo alla formula "il visibile non è riducibile al dicibile", questa non è, per usare la distinzione di Wittgenstein in *Della Certezza* (2016, §§57, 58, 468, vedi anche 2014, §§90, 251), una proposizione empirica, è piuttosto una proposizione grammaticale. Nel nostro mondo largamente dominato dalla parola è *necessario supporre* che esista qualcosa che la parola non riesce a catturare. Rimane, tuttavia, che il visibile può essere *oggetto* di esperienza solo in quanto sussunto semioticamente, cioè in quanto è pensato come un segno e mediante operazioni segniche.

I LM (e i VLM) sono casi esemplari di questa insuperabile semiotizzazione dell'esperienza, dal momento che esibiscono in modo affatto evidente l'impossibilità di un rapporto diretto, non mediato semioticamente, con il mondo. Se già 'semplicemente' vedere il mondo è un'operazione semiotica (Cimatti, 2024), non stupirà che questo stesso vedere sia così facilmente riportabile in parole. Rimane tuttavia una tensione fra il nome e l'immagine, una tensione intrinseca in ogni operazione semiotica, una tensione che rimanda al fatto che ogni segno è allo stesso tempo determinato e indeterminato.

Bibliografia

Capone L., *Wittgenstein's remarks on the philosophy of computational linguistics. Philosophical clarification applied to natural language processing*, in "Rivista Italiana di Filosofia del Linguaggio", SFL – Rules, praxis and forms of life (2024), pp. 172-186

Capone L., *Segni e suoni: un'analisi linguistica della semantica dei codici musicali*, in "De Musica", 1(XXVII)(2023), pp. 28-48.

Capone L., *Which theory of language for deep neural networks? Speech and cognition in humans and machines*, in "Technology and Language", vol. 2 (2021), pp. 29-60
Capone L., Cimatti F., *Sensi senza significato. La competenza semantica al tempo dei "Large Language Model"*, in "Reti, saperi, linguaggi", 1 (2025), pp. 175-198
Cimatti F., *L'occhio selvaggio. Sul lascarsi vedere*, Quodlibet, Macerata 2024.
Cometa M., *La scrittura delle immagini. Letteratura e cultura visuale*, Raffaello Cortina Editore, Milano 2011
De Mauro T., *Minisemantica dei linguaggi non verbali e delle lingue*, Laterza, Roma-Bari 2024.
De Mauro, T. (1970). *Introduzione alla semantica*, Laterza, Roma-Bari.
Eco U., *Kant e l'ornitorinco*, La nave di Teseo, Milano 2016
Eco U., *Trattato di semiotica generale*, Bompiani, Milano 1975
Garroni E., *Immagine Linguaggio Figura*, Laterza, Roma-Bari 2005
Gombrich E.H., *Art and Illusion: A Study in the Psychology of Pictorial Representation* (1960); transl. *Arte e illusione: Studio sulla psicologia della rappresentazione pittorica*, Phaidon, New York 2002
Hasenmueller C., *'A picture is worth a thousand words': how we talk about images*, in "Semiotica", 73/3-4 (1989), pp. 275-300
Nichol, A.Q., *et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,* in "International Conference on Machine Learning" 2022, pp. 16784-16804.
OpenAI. *GPT-4 Technical Report,* (2023), URL = https://openai.com/research/gpt-4.
Platone, *Fedro*, Einaudi, Torino 2011
Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. *Hierarchical text-conditional image generation with clip latents* (2022), URL = https://doi.org/10.48550/ARXIV.2204.06125.
Saharia, C., *et al. Photorealistic text-to-image diffusion models with deep language understanding.* In "Advances in neural information processing systems", 35 (2022), pp. 36479-36494.
Saussure, F., *Cours de linguistique Générale* (1916); transl. *Corso di linguistica generale*, Laterza, Roma-Bari, 2015
Violi, P., *Significato ed esperienza,* Bompiani, Milano 1997
Wittgenstein, L., *On certanity* (1969); transl. *Della certezza*, Einaudi, Torino 2016.
Wittgenstein, L., *Philosophische untersuchungen* (1953); transl. *Ricerche filosofiche*, Einaudi, Torino 2014.

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221514

Reimagining AI Perception Cybernetics, Extractivism, and the DIY Turn

Samuele Sartori*

Abstract

This article explores the perceptual epistemologies underlying contemporary artificial intelligence (AI), tracing their origins to cybernetic theories of perception, learning, and feedback. While generative AIs inherit a bottom-up model of behavior formation from cybernetics, their current implementation is largely shaped by extractive logics – both in terms of data and hardware infrastructure. These logics obscure the perceptual mechanisms, training datasets, and material embodiment of AI systems, leading to multiple "black boxes". As a counterpoint, the article highlights alternative approaches found in DIY culture and performance art, such as Gordon Pask's *Musicolour*, Luca Pagan's *Multi-Node Shell*, and the community-driven AI *IAQOS* by Salvatore Iaconesi and Oriana Perisico. These projects demonstrate the possibility of ethical, situated, and relational models of machine perception and agency. Ultimately, the article advocates for a reimagining of AI as a co-constituted, materially grounded, and participatory technology, capable of fostering meaningful interactions between human and non-human agents.

Keywords

Cybernetic robotic; DIY AI; media archeology; media epistemology; AI ethics

1. *Introduction: Epistemologies and Technologies of Perception*

When did the inorganic begin to have its own form of perception? This question invites a variety of epistemological responses, each offering radically different frameworks. For instance, scholars of new materialism would argue that perceptual and intra-active modes are inherent to all matter, challenging rigid dichotomies such as sensitive vs. insensitive. These oppositions, they contend, perpetuate an anthropocentric conception of embodiment that must be rethought (Barad 2007; Bennett *et al.* 2010). From another angle, media ecology would claim that it is in the very nature of mediality to turn matter into an extension of perception. In this view, a specific class of technologies, media, function as prolongations of human sensitivity, ultimately becoming integral to it (McLuhan

* samuele.sartori@unimi.it, University of Milan, Department of Philosophy

1964; Cali 2017). By contrast, a more biological or physiological approach would find little continuity between organic perceptual modes and the binary logic underpinning today's digital devices. From this perspective, current digital systems lack the embodied, sensorimotor grounding that defines biological perception.

The diversity – and often the contradiction – of these responses stems from the fact that each discipline defines perception according to its own epistemology. As a result, continuities and discontinuities between the organic and the inorganic are inferred based on how each model conceptualizes perception. This carries the risk of falling into circular reasoning, where premises and conclusions become indistinguishable. To avoid this potential pitfall, I believe it is essential to begin with a specific family of objects, thereby narrowing the scope of the question. My analysis does not concern the inorganic in general, but rather those artifacts and devices explicitly designed to mimic perception. As will become clear throughout this study, although many perceptual technologies originate from a representational ideal – seeking to establish analogies between animal and mechanical or digital forms of sensititvity – the outcomes often deviate significantly. These technologies may enhance, diminish, or even transform the very perceptual capacities they aim to replicate.

Without resorting to simplistic forms of determinism – be it sociological or technological – I will approach the field of robotics as a media-archaeological archive, a space through which it will be possible to trace the relationships between epistemologies of perception and the production of technologies capable of embodying them. The aim will not be to offer a historical reconstruction of the key stages that led to the emergence of AI. Rather, the goal will be to expose the deep entanglement between perception and technological agency in contemporary AIs, and to show how these technologies inherit and transform certain epistemological frameworks – most notably, that of cybernetics. At the same time, the article will highlight the ethical and political limits of AI's current forms of embodied perception and agency. In response, it will propose possible alternatives drawn from the field of performative and interactive digital art, where artists develop their own AI systems as critical, embodied, and aesthetic interventions.

2. *From Environment to Learning*

We might begin this inquiry with a simple slogan: "perception means having access to an environment". Hugo Gernsback – a prolif-

ic inventor active in early 20th-century America – understood it well. Many of his creations embody this idea with striking clarity. Among them: an early prototype of virtual reality that immersed users in a dynamic environment of geometric patterns, which Gernsback claimed could enhance reasoning and concentration; a precursor to the mobile phone, using radio frequencies to allow communication among people within a shared media space; and, most relevant to this study, one of the first robotic dogs (Gernsback 2016). Advertised in 1923 in *Practical Electrics* – a Do It Yourself (DIY) electronics magazine he founded – The Electric Dog was a rudimentary machine that "follows a magnetic walking cane in any direction like an obedient servant" (Gernsback 1923, p. 496). Despite its simplicity, the robot could detect a specific external stimulus and respond accordingly. In Gernsback's design, the environment was reduced to a unique, trackable input, and the action repertoire of the robot was binary: move or remain still. In this performance, The Electric Dog embodied the logic of behaviorism. This school of physiological psychology – dominant from the late 19th to early 20th century – proposed that both animal and human actions could be explained through the stimulus-response paradigm: a given perception triggers a predictable action (Watson 2017). Gernsback's robotic dog thus instantiated a linear, reactive model of agency, structured around the basic unit of perturbation and response.

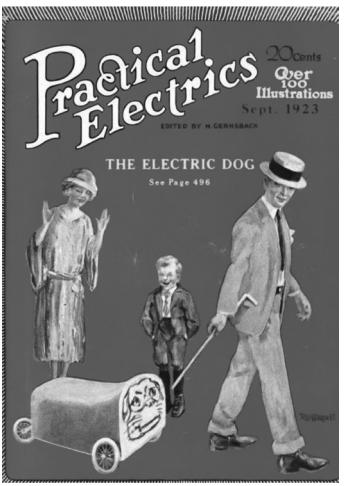


Fig. 1 – Cover image of the September 1923 issue for *Practical Electrics*, edited by Hugo Gernsback.

However, the stimulus-response model, along with the mechanistic framework that underpinned this form of behavioral physiology, soon began to falter – both theoretically and technologically. Biologists, like Jakob von Uexküll, pragmatist philosophers, notably William James and John Dewey, and phenomenologists, such as Maurice Merleau-Ponty, challenged the dualist and atomistic assumptions of behaviorism. This tradition treated stimulus and response as separate, autonomous events. They also rejected its notion of passive stimulation and its limited capacity to explain how new habits emerge. From this critical foundation, these thinkers – despite their theoretical differences – reconceptualized perception, movement, and cognition as intrinsically interrelated, co-constituted processes, deeply informed by an personal and cultural engagements with the environment. In this alternative framework, behavior and the formation of habitual actions were not the result of mechanical reactions, but rather the product of a dynamic negotiation between the organism's needs and the affordances of the environment. This relationship was corporeal, malleable, and capable of generating new goal-oriented activities, shaped and polarized by lived experience (von Uexküll 2010; James 1983; Dewey 1896; Merleau-Ponty 1967; Miglio & Sartori 2021).

It was Norbert Wiener – along with the science he founded, cybernetics – who inherited and expanded this new epistemology of the body, translating it into the domain of the non-living. Emerging in the 1940s from the interdisciplinary collaboration of mathematicians, engineers, psychologists, physiologists, and cognitive scientists, cybernetics positioned itself as the study of communication and control in animals and machines (Husbands & Holland 2008). The goal of the new science was twofold: either to transfer organic capabilities into machines (robotics), or to enable communication and control between organisms and technologies (prosthetics). Whether in the form of robotic replication or bodily integration, cybernetics introduced an embodied epistemology centered on the concepts of perception, homeostatic balance, and feedback loops. Perception was rethought through an interpretation of information theory, as developed by Claude Shannon and Wiener himself. Information theory aimed to mathematically describe and quantify how a message (or input) is encoded, transmitted, received, and decoded into an output. Within this framework, perception is redefined as the organism's capacity to extract and process environmental signals – not passively, but through active engagement. Perception becomes both an attentional and motor act: a posture of listening and a preparation for

response. At the same time, the biological body is also a source of output, feeding information back into the environment. This bidirectional flow creates a feedback loop, through which action and cognition are shaped by interaction with the environment, and vice versa. The ultimate aim of this interplay is self-organization, oriented toward maintaining homeostasis – a dynamic equilibrium that preserves life and environmental coherence. Another foundational principle of cybernetics was its bottom-up model of cognition: thoughts and beliefs do not precede experience, but are semi-randomly constituted and continually modifiable by it. Through the repeated association of perceptions and actions, patterns of thought emerge – or, in cybernetic vocabulary, behavioral patterns. In short, the aim of cybernetic research was not to treat perception, homeostasis, and feedback as isolated phenomena, but to demonstrate how their interaction generates new behaviors, and how these behaviors solidify into habitual but still transformable structures of action (Wiener 1948).

As previously mentioned, cybernetics was not merely a conceptual or theoretical perspective – it also gave rise to some of the earliest technological prototypes that sought to robotically embody its epistemology. One notable example is the Homeostat, a device designed by psychiatrist William Ross Ashby to model adaptive behavior through mechanical means. The Homeostat was a machine capable of random self-adjustment in order to maintain internal equilibrium. It consisted of four interconnected units, each equipped with a pivoting magnet representing a system variable. Each unit's output influenced the others, with the magnet's torque proportional to the total incoming electrical current. The output was tied to the magnet's deviation from its central position, while the behavior of the system was shaped by parameters such as potentiometer settings and commutator configurations. The system employed primary feedback via electrical interactions, and a secondary feedback loop that activated when the deviation of a magnet exceeded its limit. In such cases, the machine would randomly adjust internal parameters until a stable state was restored – at which point the secondary feedback would shut off. In short, despite its mechanical simplicity, the Homeostat exhibited a primitive form of self-perception, prompting action whenever its equilibrium was disrupted. Rather than resisting instability, the machine treated error and imbalance as opportunities – a moment in which mechanical matter could experiment with variation in order to regain homeostasis. Ashby's Homeostat thus represents an early attempt to simulate a kind

of autonomous learning: a system that, through trial and error, seeks internal solutions to external perturbations using only its own resources (Ashby 1945; Asaro 2008).

Cybernetics shared its embodied epistemology with the emerging connectionist wave that would lead to the development of one of the earliest neural networks: the Perceptron, conceived primarily for military applications by psychologist Frank Rosenblatt and Marshall Yovits of the U.S. Office of Naval Research (Palfreman & Swade 1991, pp. 136-167). Before introducing the Perceptron in detail, it is important to note that, in the 1950s, the earliest AI models attempted to simulate the behavior of a single biological neuron. This principle formed the structural basis of Rosenblatt's machine. The Mark I Perceptron was a simple, three-layer neural network consisting of sensory units (S-units), associative units (A-units), and response units (R-units). Its input layer – a 20×20-pixel camera equipped with 400 photoreceptors – was connected either randomly or topologically to 512 associative units with fixed synaptic weights. These associative units, in turn, were linked to eight response units with adjustable weights, modulated via analog potentiometers that could also be manually fine-tuned (Minsky & Papert 1969).

The Perceptron's primary application was in what we now call pattern recognition. Through a training process based on error correction, the Perceptron was designed to construct probabilistic patterns that would allow it to decode visual inputs. Its combined hardware and software enabled it to recognize numbers and letters, as well as to approximate binary gender classification (male/female) based on facial photographs. As Rosenblatt himself emphasized, "the proposed system depends on probabilistic rather than deterministic principles for its operation" (Rosenblatt 1957, p. 2). These probabilities were not fixed, but evolved through the machine's experience, adjusting with each new input-output cycle. For this reason, the Perceptron is widely considered one of the first bottom-up AI models. Rather than relying on predefined rules, the model learned from experience, starting with raw sensory data and gradually building toward classification by modifying the internal weights of its network based on performance feedback.

In summary, both of these machines embodied the dream – and to some extent the capability – of adapting and modulating actions based on a semi-random (Homeostat) or statistical (Perceptron) reprocessing of technologically mediated perception. But what remains today of this epistemology in contemporary AI systems? And how has it been transformed over time?

3. *Between Cybernetics and Contemporary AI: Continuities and Discontinuities*

The principles of perception, homeostatic balance, and feedback loops – which underpin the acquisition of new capabilities through semi-random elaboration and the consolidation of habits via statistical repetition – are also central to deep learning. These software systems employ multi-layered neural networks that, while far more complex and hierarchical than their predecessors, still trace their lineage back to the architecture of Rosenblatt and Yovits's Perceptron (Kurenkov 2020; Goel 2021). At the same time, however, contemporary AI systems operate as agents that are not necessarily robotic, but increasingly and primarily digital entities capable of interacting with dynamic, datafied environments.

Take, for example, Albert, created by AI Warehouse. In the software, Albert is learning how to escape from five different rooms. The agent's movements are governed and modified by five layers: the first layer processes the inputs – information Albert receives before acting, such as limb positions and velocities, representing a dynamic, homeostatic balance that the agent must partially maintain to perform actions. The final layer dictates the actions to be taken, while the three intermediate hidden layers perform the complex calculations needed to transform inputs into actions. Unlike traditional approaches relying on motion capture of real humans, AI Warehouse's software employs deep reinforcement learning exclusively to teach Albert to walk in a human-like manner. Each room presents Albert with the same set of possible actions, but the reward-punishment system varies across environments. What is particularly fascinating is how the AI adapts its movements by perceiving and acting within new spaces. For instance, in the first room, Albert can crawl to reach the goal, but the floor in the second room prohibits this action, forcing Albert to modify its walking behavior accordingly. The challenges encountered in each room progressively lead Albert to build, strengthen, or transform certain behavioral habits. Specifically, performances that statistically prove effective for interacting with the situated environment are reinforced, while less efficient ones are semi-randomly altered.[1] This internal feedback

[1] The alteration is semi-random, as Albert has constraints imposed by the design of his digital body. In particular, his creator writes: "For each of Albert's limbs I've given him (as an input) the position, velocity, angular velocity, contacts (if it's touching the ground, wall or obstacle) and the strength applied to it. I've also given him the distance from each foot to the ground, direction of the closest target, the direction his body's moving, his body's velocity, the distance from his chest to his feet and the amount of time one foot has been

loop, grounded in experience, ultimately enables Albert to develop an increasingly stable and human-like gait.

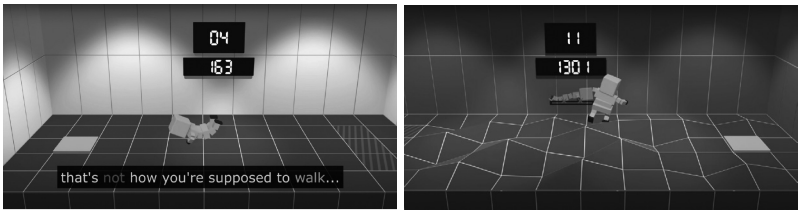


Fig. 2-3 – Movement frames of Albert: on the left, Level 1; on the right, Level 2. The orange dashed rectangle marks the starting point, while the green square indicates the goal. The numbers at the top show the remaining seconds to complete the movement (each attempt allows the agent 1 minute), and the numbers at the bottom indicate the attempt count (these frames were captured during the successful attempt in which Albert completed the level).

Deep reinforcement learning, which underlies the functioning of agents like Albert, is a subfield of machine learning that combines the principles of reinforcement learning and deep learning. Reinforcement learning focuses on optimal control – it studies how an intelligent agent should take actions within a dynamic environment to maximize a reward signal. Deep learning, on the other hand, is a branch of machine learning that uses multilayered artificial neural networks to perform complex tasks such as classification, regression, and representation learning. These models consist of multiple layers of interconnected neurons, where each layer progressively extracts higher-level features from the input data. The term "deep" refers to the presence of many layers – ranging from a few to hundreds or even thousands – allowing the network to learn highly intricate patterns. It is primarily deep learning systems, working through trial and error, that enable the operation and advancement of the most widespread AI applications today, such as ChatGPT, Google AI, and recommendation systems on YouTube, Instagram, and TikTok, but they also power game-playing simulators like AlphaZero chess bot and driving simulators like Gran Turismo Sophy.

One could delve deeply into the software dimension of generative AI (Mattingly & Cibralic 2025), but such a description, no matter how detailed, would remain disembodied – failing to

in front of the other. As for his actions, we allow Albert to control each body part's rotation and strength (with some limitations so his arm can't bend backwards, for example)".

consider both the data and the hardware body essential for developing these technologies (Ernst 2021). Generative AI requires vast amounts of data, which effectively form a digital environment. Only through interacting with this environment can generative AI transform and optimize itself for the purposes it was designed to serve. However, it is crucial to emphasize that this environment is not neutral. Numerous studies have revealed racist and sexist biases embedded within AI systems, largely stemming from the data archives on which these models are trained (Zuboff 2019; Broussard 2023; Pasquinelli 2023). Moreover, this environment is not directly manipulable by humans – it consists of an immense sea of bits and information that far exceed our cognitive capacities. The demand for massive data sets has driven AI companies to gather information indiscriminately, often disregarding its origins. This extractive strategy frequently involves collecting data without the knowledge or consent of the users who generated it (Stalter 2018; Casilli 2025).

Contemporary hardware also reflects the logic of extractivism, most notably through the ecological devastation caused by the extraction of raw materials – particularly rare metals – required for constructing the physical infrastructure of AI systems (Crawford 2021; Pitron 2021). Furthermore, the body of AI is highly complex and difficult to pinpoint because it lacks a single, fixed location: our personal computers are part of it, as are phones, VR headset and smartwatches, along with the hardware that actually processes the data. This creates a crucial distinction from the cybernetic model: while robotic perception in cybernetic thought is localized within a physical environment, the digital perception of contemporary AIs is distributed across a vast array of devices that generate the data, forming a rich digital world for AIs. Yet, this environment is scarcely locatable within any single physical space – or even across multiple physical spaces.

In summary, we are witnessing the rise of multiple "black boxes": we do not truly understand the processes occurring within the various layers of neural networks during AI generative tasks (that is, the decoding of perception); we lack knowledge about the data used to train these AIs (the objects of digital perception); and we remain unaware of the AI's body (the material substrate capable of perceiving). While cybernetic epistemology still provides a theoretical framework for understanding the relationship between perception and learning in contemporary AI, the extracrtive model – problematic from both ethical and political standpoints – dominates the practical functioning of these technologies and it shapes their specific operativity.

4. *An Alternative to Extractive Perception*

In what domains can one identify models that counter the extractivist approach to AI? Once again, both from a technological and epistemological perspective, cybernetics offers valuable insights. First, it is important to recognize that Norbert Wiener was already aware of the potential dangers of subjugation and dependency on technology. In his co-constitutive, bottom-up model, humans are called to reorganize and relearn how to think with and through cybernetic tools. However, to guide these devices rather than become enslaved by them, a reflective, material, and computational understanding of the body and functioning of contemporary technologies is essential. This kind of literacy – or "grammatization", as Bernard Stiegler would say (2015) – is necessary not only to shed light on the black boxes that constitute AI, but also to transform and manipulate these devices. Critical awareness, therefore, is not just reflective knowledge; it is, above all, a practical capacity to act and to redirect the purposes for which certain technologies were originally designed (Wiener 1950).

A fruitful expression of the possible connections between theory and practice in Artificial Intelligence technologies can be found in the experiments of the DIY culture. Gernsback, with his magazines teaching how to dismantle and modify early household electronic devices, along with many of the most brilliant cyberneticians, can be considered precursors of this hybrid current – one capable of reimagining the worlds of engineering, art, and reuse. A particularly prominent and pioneering figure in this practice was Gordon Pask. His *Musicolour*, developed in the 1950s in a small room as a side project during his medical studies, was a cybernetic device capable of interacting with musicians by dynamically modulating stage lighting in semi-structured, automated ways responding to live music. This created a performative feedback loop between human and machine – a transaction of perception and action. Materially, the music was converted into an electrical signal by a microphone, which, inside *Musicolour*, passed through a set of filters sensitive to different frequencies, beats, and so on, with the outputs controlling different lights. One might imagine the highest-frequency filter activating a bank of red lights, the next-highest triggering blue lights, and so forth. The concept was simple, except for one key detail: the internal parameters of *Musicolour*'s circuitry were not fixed. Analogous to biological neurons, banks of lights would only activate if the output from the relevant filter exceeded a certain threshold – and these thresholds varied

over time as charges built up on capacitors, influenced by the ongoing performance and the machine's prior behavior (Pickering 2010, pp. 309-321).

In summary, *Musicolour* was an AI with generative capabilities and situated perception, able to modify its actions not only based on interactions but also in response to exposure to particular engagements. As Pask himself wrote a few years later:

> By that time it was clear that the interesting thing about *Musicolour* was not synaesthesia but the learning capability of the machine. Given a suitable design and a happy choice of visual vocabulary, the performer (being influenced by the visual display) could become involved in a close participant interaction with the system. He trained the machine and it played a game with him. In this sense, the system acted as an extension of the performer with which he could cooperate to achieve effects that he could not achieve on his own. (1971, p. 78)

This is not about finding a necessary reconciliation in the possible relationships between humans and technology, but rather about ensuring an exploratory space – uncertain and open to experimentation. Moreover, it is worth noting that this intelligent structure led many artists who used it to lose track of time, engaging for hours in play with the *Musicolour*. These were not mere repetitions but variations. The stimulus does not elicit a fixed response; instead, the machine reacted to sonic sensations with modulations that generated clashes and synesthetic experiences within a poietic environment of silent shaping – a medial space existing beyond the dichotomies of efficiency and inefficiency.

The case of *Musicolour* is by no means isolated, nor is it the only technology in the performance world to embrace a development of AI grounded in the perceptive-agentive corporeality of cybernetics, rejecting the extractive model. One could continue by analyzing the works showcased in the 1968 exhibition Cybernetic Serendipity at the Institute of Contemporary Arts in London, which gathered various projects and reinterpretations of Wiener's young science. From there, we arrive at contemporary experiments such as those by Luca Pagan, a sound designer and performance artist who builds both hardware and software for wearable technologies. In particular, one of these exoskeletons, *Multi-Node Shell*, is driven by a generative algorithm capable of transforming the performer's body movements into non-standardized modulations of sound, continuously creating new correspondences between gesture and music. While fascinating, a thorough historical reconstruction of the intersections between AI, DIY culture, and performance art would be too extensive to cover here, given the vast amount of material

available (Diederichsen & Etxeberria 2021). Before concluding this section, however, I want to consider one final, central case study that illustrates how these DIY technologies can transition from the world of art to everyday life.

When creating *IAQOS*, Salvatore Iaconesi and Oriana Perisico aimed to gift the Torpignattara neighborhood in Rome an AI designed for its inhabitants. Unlike most Artificial Intelligences, *IAQOS* does not operate extractively – accumulating data to profile people and steer their behavior – but relationally – connecting people so they can become active agents of social transformation. The AI learns from the community it lives with and grows alongside it. *IAQOS* is therefore designed from the ground up, through workshops where residents discuss the identity of their neighborhood, their unmet needs, and visions for the future. For example, *IAQOS* spends time at the Pisacane school, where children do their homework together with it, and elders share stories to pass down. The information *IAQOS* receives shapes its behavior and, as both a mirror and part of the community, its actions can serve as a social and local alarm bell: for instance, if *IAQOS* exhibits racist behavior, it signals a problem in the neighborhood that residents can recognize and address collectively. Through these varied interactions, the AI prompts citizens to reflect on the central role data plays in their daily lives, fostering awareness of technology's impact and the importance of understanding it – not to be passively subjected to its effects, but to actively harness its opportunities (Iaconesi & Perisico 2021).

5. *Conclusions: Possible Futures for the Perception of Artificial Intelligence*

This article has outlined a potential media-archaeological and epistemological history of AI's operativity. The emphasis on perception as a co-constitutive factor in shaping behaviors and habits, evident in contemporary generative AI, is a legacy of cybernetics – particularly the bottom-up model it developed to describe the emergence of human and technological consciousness within experience. Beyond acknowledging this origin, the paper has also examined the discontinuities that contemporary AI models embody in relation to their epistemological roots. These divergences are most evident in the adoption of an extractive paradigm, especially regarding ethical and political dimensions. While cybernetics recognized the social necessity – not merely a specialist concern – of critical

and pragmatic education about technologies, extractivism has led to the proliferation of black boxes, obscuring understanding and control for the user.

To move beyond the extractive model and the black boxes it creates, I have proposed turning to the DIY culture connected to performance art, which produces its own forms of AI. Projects like *Musicolour*, *Multi-Node Shell*, and *IAQOS* subvert the neo-capitalist logic underpinning the most widespread AI systems. While extractivism devastates ecological balances by exploiting raw materials, DIY culture experiments often rely on media and tools made from reused or repurposed components. On one hand, this is an act of *détournement* – transforming the original purposes for which certain technological components were designed. On the other, it is a conscious acknowledgment of how we can actively manipulate the hardware elements of the technologies surrounding us. More importantly for the theme of perception, these case studies rethink the origins of data, their distribution, and the purposes behind their processing. Whereas extractivism indiscriminately harvests data without consent to provide stimuli and create the digital environments in which AI is trained, these artistic projects prioritize not just consent but the conscious choice to establish a perceptual communication between the human and technological bodies. Finally, these experiments suggest that an ethical exchange – deliberate, sought after, and poietic – between different forms of perception is not merely a hopeful ideal but a materially and technologically viable alternative reality. This represents a new opportunity to reimagine AI's operativity, which I hope will move from the realm of art into everyday life.

Bibliography

Asaro M., 'From Mechanisms of Adaptation to Intelligence Amplifiers: The Philosophy of W. Ross Ashby', in P. Husbands, O. Holland, M. Wheeler (eds.), *The Mechanical Mind in History*, MIT Press, Cambridge 2008, pp. 149-184.

Ashby W.R., *The Physical Origin of Adaptation by Trial and Error*, in "Journal of General Psychology", 32 (1945), pp. 13-25.

Barad K., *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, Durham, Duke University Press 2007.

Bennett J., *et al.* (eds), *New Materialisms: Ontology, Agency, and Politics*, Durham, Duke University Press 2010.

Broussard M., *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*, MIT Press, Cambridge 2023.
Cali D*., Mapping Media Ecology*, New York, Peter Lang Press 2017.
Casilli A.A., *En Attendant les Robots: Enquête sur le Travail du Clic* (2019); trans., *Waiting for Robots: The Hired Hands of Automation*, University of Chicago Press, London 2025.
Crawford K., *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence,* Yale University Press, Yale 2021.
Dewey J., *The Reflex Arc Concept in Psychology*, in "Psychological Review", 3 (1896), pp. 357-370.
Diederichsen D., Etxeberria O. (eds.), *Cybernetics of the Poor*, Sternberg Press, Berlin 2021.
Ernst W., *Technológos in Being: Radical Media Archaeology & the Computational Machine*, Bloomsbury, New York 2021.
Gernsback H., *The Electric Dog,* in "Experimental Electrics", September 1923, pp. 496-497.
Gernsback H., *The Perversity of Things: Hugo Gernsback on Media, Tinkering, and Scientifiction*, Minneapolis, University of Minnesota Press 2016.
Goel A.K., *Looking Back, Looking Ahead: Symbolic Versus Connectionist AI*, in "AI Magazine", 42 (2021), pp. 83-85.
Husbands P., Holland O., 'The Ratio Club: A Hub of British Cybernetics', in P. Husbands, O. Holland, M. Wheeler (eds), *The Mechanical Mind in History,* MIT Press, Cambridge 2008, pp. 91-148.
Iconesi S., Perisico O., *When my Child is AI: Learning and Experiencing through AI outside the School, the experiences of a community AI*, in "QTimes: Jorunal of Education, Technology and Social Studies", 13/1 (2021), pp. 174-172
James W., *Principles of Psychology* (1890), Harvard University, Harvard Press 1983.
Kurenkov A., *A Brief History of Neural Nets and Deep Learning*, in "Skynet Today: Putting AI News In Perspective", 27 September 2020, skynettoday.com [accessed 7 January 2025].
Mattingly J., Cibralic B., *Machine Agency*, MIT Press, Cambridge 2025.
McLuhan M., *Understanding Media: The Extensions of Man*, McGraw Hill, New York 1964.
Merleau-Ponty M., *La Structure du Comportement* (1942); transl. *The Structure of Behavior*, Beacon Press, Boston 1967.
Miglio N., Sartori S., *Perceptual and Bodily Habits: Towards a Dialogue Between Phenomenology and Somaesthetics*, in "The Journal of Somaesthetics", 7/1 (2021), pp. 29-44.

Minsky M., Papert S., *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge 1969.
Palfreman J., Swade D., *Dream Machine: Exploring the Computer Age*, BBC Books, London 1991, pp 136-167.
Pask G., 'A Comment, a Case History and a Plan', in J. Reichardt (ed.), *Cybernetics, Art, and Ideas*, York Graphics Society, New York 1971, pp. 76-99
Pasquinelli M., *The Eye of the Master: A Social History of Artificial Intelligence*, Verso Books, London 2023.
Pickering A., *The cybernetic brain: Sketches of another future*, University of Chicago Press, Chicago 2010.
Pitron G., *La Guerre des Métaux Rares: La Face Cachée de la Transition Énergétique et Numérique* (2018); trans. *The Rare Metals War: The Dark Side of Clean Energy and Digital Technologies*, Scribe, London 2021.
Rosenblatt R., *The Perceptron. A Perceiving and Recognizing Automaton*, Cornell Aeronautical Laboratory Report No. 85-450-1, Buffalo 1957.
Stalder F., *Kultur der Digitalität* (2016); trans. *The Digital Condition*, Publisher Polity Press, Cambridge 2018.
Stiegler B., *De la Misère Symbolique: La Catastrophe du Sensible*; trans. *Symbolic Misery: The Katastrophe of the Sensible*, Polity Press, Cambridge 2015.
von Uexküll J., *Streifzüge durch die Umwelten von Tieren und Menschen: Ein Bilderbuch unsichtbarer Welten* (1934); transl. *A Foray into the Worlds of Animals and Humans (with A Theory of Meaning)*, University of Minnesota Press, Minneapolis 2010.
Watson J.B., *Behaviorism* (1924), Routledge, London 2017.
Wiener N., *Cybernetics: Or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge 1948.
Wiener N., *The Human Use of Human Beings: Cybernetics and Society*, Houghton Mifflin Harcourt, Boston 1950.
Zuboff S., *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile Books, London 2019.

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221515

What Effect Does Your Nonhuman Touch Have on Me? An Aesthetic Approach to Human-Robot Interaction

Chiara Cappelletto[*], Giulio Galimberti [**]

To S.B., tireless breaker of things, forever blaming them for what she makes them do.

Abstract

This paper proposes an aesthetic approach to Human-Robot Interaction (HRI), focusing on the bodily dimension of human experience with social robots in tactile encounters. We resist cognitive-centered models, and value robotic materiality and affordances – particularly heat. Drawing from an experimental study with the NAO robot, we argue that thermoception, often neglected in HRI research, plays a critical role in shaping our emotional engagement with and acceptance of robotic agents. Their bodies, we thereby contend, are not just neutral tools but relational partners capable of altering our expectations and challenging our own bodily habits thanks to their material heterogeneity.

Keywords

Social robots; haptics; thermal HRI; material performative agency; bodily transparency.

1. *Too-Intelligent Robots*

Robotic systems are primarily conceived of as tools that replace or enhance our cognitive and bodily behaviors oriented toward a useful purpose. Their *raison d'être* is to satisfy objectives deemed advantageous. Advertisements for little vacuuming robots that clean the house while the family relaxes serve as an example of the widespread belief that intelligent machines must adapt to our needs. According to this logic, robots are simply subordinate and

[*] Department of Philosophy "Piero Martinetti", University of Milan, Milano, Italy;
PIS – Performing Identities Studies Lab, Department of Philosophy "Piero Martinetti", University of Milan, Milano, Italy;
PIS unit, CRC EXT, Department of Philosophy "Piero Martinetti", University of Milan, Milano, Italy
[**] Department of Philosophy "Piero Martinetti", University of Milan, Milano, Italy;
PIS – Performing Identities Studies Lab, Department of Philosophy "Piero Martinetti", University of Milan, Milano, Italy;
PIS unit, PHILAB, Department of Philosophy "Piero Martinetti", University of Milan, Milan, Italy

efficient things. The value of human-robot interaction (HRI) is thus discussed in terms of the level achieved by these nonhumans as to the comprehensibility of the information produced (Wang *et al.* 2025), their fluency in task execution (Ali, Green & Iqbal 2024), or the acceptability of their behavior in relation to the judgment of "non-artificial" agents (Kraus *et al.* 2022).

The point we wish to challenge here is the idea that semi-autonomous robots must act as we ourselves would if they are to meet our needs. This argument is deep-rooted. The epistemic forebear of all artificial intelligence is indeed the Turing machine, which in 1936 established our epistemic horizon of expectation: the simulation of human abilities, at least to the point of being indistinguishable from them, if not surpassing them. In our view, this authoritative tradition is fueled by an implicit theoretical assumption: that HRI is nothing but a contest between brainocentric agents, whose epic moment is represented by the chess match between Garry Kasparov, the Russian world champion who defeated Deep Blue in 1996, only to be beaten by IBM's supercomputer one year later.

This agonism between rivals, defined by their cognitive faculties – independent of any emotional dimension – is conservative both heuristically and epistemologically, and, as we will later see, socially as well. The idea that, to interact effectively, we need but train ourselves to use a machine that, in turn, is predisposed to be as user-friendly as possible, is reductive for understanding the nature not only of things but of ourselves. Its first result is a technophobia whereby we worry that the machine will rebel, just as we would. The 1927 film *Metropolis* by Fritz Lang tells the story masterfully. It is an unresolved fear.

The theoretical assumptions underlying the development of intelligent mechanical systems reproduce a logocentric epistemology in which HRI is primarily modeled as informational exchange between rational agents (Suchman 2007), where "reason" means calculation and prediction, neglecting the emotional capabilities of our embodied mind (Cavarero 2003; Gallese 2007). Framing the question in rationalistic terms not only affects how we design and treat robots but also reinforces a specific understanding of our own nature: if they are rational – that is, disembodied – because they replicate us, then we, as the ones replicated, are disembodied too. Mentalizing them means accepting as the model of an average human someone who has needs but no desires, is responsive to stimuli but insensitive, efficient but unambitious, without gender, age, ethnicity, social or educational characteristics, and free of flaws. Indeed, a robot.

Yet, a simple return to science fiction – from *I, Robot* by Isaac Asimov (1950) to *Klara and the Sun* by Kazuo Ishiguro (2021), from *2001: A Space Odyssey* by Stanley Kubrick (1968) to *Her* by Spike Jonze (2013) – might help us challenge this narrative and prompt us to ask if the rationalistic privilege of artificial intelligence is not in fact hallucinatory: don't robots show that what is actually human is our capacity for play, rebellion, and error? Doesn't HAL – a pure interface – start becoming like us the moment it stops calculating correctly and begins to aspire, object, kill, and fear? Above all: doesn't Dave, the astronaut who deactivates it, save himself by opening a hatch whose handle could only be turned manually?

We do not mean to suggest a technologically inflected alternative between (mortal) sensitivity and (superhuman) intellect, but to highlight the two defining behaviors of these artificial agents, based on their material specificity. On the one hand, we have "smart homes," in which computers control the lights and temperature, raise the shutters, and set cooking timers, thus functioning as intelligent and interactive systems, with their very body consisting of our entire house (Chan *et al.* 2008; Anderson & Anderson 2010; Telakivi 2022;). On the other, we increasingly deal in our daily lives with physical robots whose primary role is no longer task execution and that come into direct contact with us. In healthcare (Aymerich-Franch & Ferrer 2023), educational (Ceha *et al.* 2022), and entertainment contexts (Frieske *et al.* 2024) we rely on robots that are materially present in our surroundings, nearby, at hand. The relevant feature, in such cases, is their ability to be accepted by their living partners, whose subjective experience they affect and who are expected to want them around,[1] even in their most cherished spaces and in situations of vulnerability – for instance in assisted-living and caregiving contexts. In this second case, which is the one we will focus on, specific bodily habits are needed for both agents, potentially leading to a revision of the very principle of delegation, whereby the robot's ability to act on our behalf in sensible tasks is acknowledged (Latour 1992).

2. *Social Robots and Human Bodies*

NAO, PEPPER, Milo, Kaspar, and ZORA are some names for so-called social robots, also known as "so-bots" (Mobed *et*

[1] In the Usability, Social acceptance, User experience, and Societal impact (USUS) model proposed by Weiss *et al.* (2011), social acceptability is defined as an individual's willingness to interact with a robot in an everyday social environment.

al. 2024), already available on the market. They most often consist of materially extended hardware with which people engage in situations of physical proximity, if not direct contact. In both private and public sectors, they are procured to populate hospitals, schools, hotels, shopping centers, and nursing homes and interact with those who inhabit these spaces. Their design aims to foster an HRI in which the aforementioned imitative criterion extends beyond mimicry of our cognitive abilities (Ventre-Dominey *et al.* 2019) to include standards such as appropriateness and credibility of physical behaviors and adaptability in nonverbal communication through gestures, expressions, and the management of proper distance (Fong, Nourbakhsh & Dautenhahn 2003). Social robotics, therefore, focuses on intelligent systems built to collaborate with human beings – understood as relational bodies – within social, community, and intimate contexts (DeFalco & Dolezal 2023). It thus offers a way out of the rationalistic hallucination described above.

The minimum criteria for qualifying an intelligent mechanical system as "social" are neither clearly defined nor consistent in the literature (Jackson & Williams 2021). In many accounts, the degree to which it visually resembles us seems to be a key requirement for ensuring its sociality (Krägeloh *et al.* 2019). But the extent to which this effect depends on the morphological similarity of the mechanical body to ours and the extent to which it derives from other features beyond visual appearance is a highly debated question (MacDorman & Ishiguro 2006; Kätsyri *et al.* 2015). Moreover, in the case of social robots, humanization seems to produce an ontological leap: they are experienced as "social others" capable of sustaining interactions comparable to those that occur between people (Jones 2017).

Since the mid-2000s, some designers have argued that robots become social not based on how familiar they look but on how their living partners rate the quality of the encounter in real time (Hegel *et al.* 2009). The judgment entailed is not purely cognitive but includes experiential and emotional dimensions, such as the feeling of being understood, the perceived motor and expressive coherence of the robot's behavior, its responsiveness, and its ability to generate trust or empathy. The emotional responses of human subjects are indeed increasingly central to the design of today's social robots. As Ericka Johnson rightly points out, "it is our emotions that are at stake in the robot-emotion relation" (2025, 1). Robots' social character, then, is an interactive condition that emerges from how we evaluate the experience of dealing with them. Redefining it as a

function of the encounter's impact on the partner means focusing on the construction of sociality as an emergent property of the engagement itself (Alač 2016).

If that is the case, robots' human-like somatic structure and their ability to replicate and/or enhance our cognitive and bodily skills no longer seem to be a sufficient or even necessary condition for them to socially engage with us. Indeed, we tend to respond favorably to mechanical partners that respect social etiquette, adapt their behavior to evolving situations, and act nicely. In this light, characterizing a robot as social means not only questioning the design choices privileged in building the artificial agent but also questioning the aesthetic conditions that make its acceptance possible – or problematic.

This becomes especially clear in situations, ranging from adolescent education to home care for the elderly, where these technologies manage complex dynamics. For example, an artificial agent in a nursing home must avoid intrusive gestures, unexpected movements, and sudden approaches and instead ensure clear and predictable motor patterns in order to be well-received by older individuals (Anderson & Anderson 2010). Similarly, a robot used as a teaching assistant must know how to stimulate student participation by capturing attention through multimodal strategies such as voice modulation, body orientation, or deictic gestures – while avoiding insistent or repetitive behaviors that could be perceived as invasive or annoying (Belpaeme *et al.* 2018).

However, this argument leads to the conclusion that robots have no actual social agency of their own, since it is humans who determine if their relational capacity is present and satisfactory. If they let us down, we can always turn them off. Are they not, after all, subordinate and performing things? This long relational detour, in the end, brings us back to the very place from where we started.

And yet, nonhuman others are increasingly required to operate in contexts where etiquette or predictability is challenged by the emotional complexity of the relationship in which they are implemented. From caregiving to companionship to mental health support, the partners often consist of individuals who, by definition, break with the idea of a statistically modeled "average" brain/body system. The evolution of social robotics thus requires us to rethink HRI not as a contest between disembodied intelligences, one of which is potentially embedded in well-trained and/or anthropomorphic mechanical structures, but as a process of relational co-construction, in which bodily, emotional, and experiential dimensions take precedence over instrumental ones and even reshape them.

3. *Tactile Feedback and New Possibilities for HRI*

Touch stands out as the sense most directly involved in interpersonal situations that prompts the co-emergence of affects, perceptions, expectations, and fears (Puig de la Bellacasa 2009; Tsirka, Velentza & Fachantidis 2024). Touch, as the relational sense *par excellence* (Merleau-Ponty 1945; Puig de la Bellacasa 2017), becomes the privileged channel for exploring new forms of coexistence and collaboration with nonliving agents, whose material constitution is aesthetically heterogeneous owing to their texture, shape, elasticity, size, humidity and weight. They produce feedback effects on the perceptual, semantic, and lived potential of tactility, which is a disposition to recognize one's own and others' bodies through the skin (Merleau-Ponty 1945). Human-robot tactile interaction (Tactile HRI) thus constitutes a growing research field within social robotics, aiming to understand how to engineer encounters that people consistently experience as pleasant, welcoming, and reassuring.

If HRI is configured as a space of negotiation in which human agents assess how much their counterpart's behavior aligns with the habits of interpersonal social life, do we expect a robot tasked with helping us up the stairs to know how to modulate the pressure of its grip the way a person would? To adjust its touch according to our sensitivity, eventually avoiding overstimulation, as would be important for some individuals with autism spectrum disorder, who register physical contact as a source of emotional stress? To have a gestural grammar like that of our bodies, which are made of bones, skin, muscles, and joints? And above all, what effect does its "sensitive know-how" have on each of us? It is once again our horizon of expectation that is called into question, here in a primarily corporeal sense, since perception is a form of first-person exploration of the environment through which we encounter material artifacts that invite us to use them in certain ways. This has important design implications, on the one hand, and heuristic and epistemic ones, on the other.

It is well-known that if tactile intimacy with a robot is deemed similar to that with a human, the exchange will be experienced positively, increasing parameters such as social presence, trust, and perceived friendship (Nie *et al.* 2012; Park & Lee 2014). However, can we be certain that a strong similarity to human touch will not reactivate that "uncanny" effect already well-documented in the case of visual resemblance, where design premises end up contradicting their own results? Masahiro Mori already showed

in the 1970s that when the appearance of a nonliving body (here a robot) approximates our features almost completely, we react with strong disbelief (Mori 1970). Does the same apply to what our skin perceives?

The term "uncanny valley of haptics" (Berger *et al.* 2018) has already been used to refer to the threshold beyond which very realistic tactile feedback can generate a sense of dissonance and reduce the degree of acceptance in HRI.[2] Similarly, Conor McGinn and Dylan Dooley (2020) showed that the texture and stiffness of robotic surfaces influence our evaluation indexes for social presence, likability, and perceived safety and that higher levels of tactile realism do not always result in more "effective" interaction. A high degree of human-likeness can thus generate discomfort or rejection. The ambiguities and indeterminacies of tactile contacts between robots and people disprove the simplistic imitative assumption whereby anthropomorphizing the material eccentricity of the artificial agent effectively neutralizes it.

Insisting on such a camouflage risks, on the one hand, to reiterate a potential uncanny effect caused by an ambitious yet imperfect and therefore fragile resemblance, and on the other, to impoverish the agentive possibilities that stem precisely from the material specificities of robotic bodies. They, for example, could maintain prolonged physical contact without tiring, getting distracted, or feeling discomfort. This kind of stable, continuous, reliable attention is not typical of us but could be perceived as reassuring and function well in therapeutic contexts. In this sense, the material heterogeneity of the robotic body can become a resource for countering the imperative to humanize the nonhuman.

In this second perspective, the robot is no longer a flawed copy of us with awkward, slightly ridiculous gestures and a hard surface, or an upgraded version that vies with us on a cognitive and operational level, but rather an agent that brings forth a specific mode of presence linked to its own gestural and relational syntax. By valorizing its bodily unfamiliarity, we mean to suggest that its design needs not to be merely modeled on our standard bodily schemes and neurobiological processes, which would in turn reduce our physicality to repetitive and identical motor skills, but instead offers a chance to expand our capacities for engagement, both with intelligent artifacts and, potentially, with our fellow humans. Wherever there is matter, there are opportunities

[2] Affective haptics – the field of research focused on designing devices capable of influencing a subject's emotional state through the sense of touch (Paterson 2023) – is becoming an essential component in HRI studies.

for a sensory interplay.

A new aesthetic experience emerges, in which the robot, whose affordances contravene our habitual expectations, exerts a specific form of "material performative agency" whereby it interferes with bodily behaviors previously familiar to us, to the extent that it disrupts our bodily transparency (Calbi & Cappelletto 2024). Material artifacts, as nonliving agents, indeed have the capacity to intervene in the brain/body system of the people with whom they interact, breaking our pre-reflective and taken-for-granted bodily experience (Leder 1990; Damasio 1994; Blanke 2012; Damasio & Carvalho 2013) and disrupting our perceptual habits. We can then coevolve with the emergence of mechanical systems that mediate touch and generate new practices of contact. The design and engineering behind recent tactile interfaces (e.g., electro-tactile feedback systems or pressure sensors and actuators embedded in robotic "skin") establish new paths for articulating the possibilities inherent in this aesthetic domain, which is increasingly relevant, not only in the design of social robots, but also in studies of the human/nonhuman relation (Sartori 2026). What if, then, our horizon of expectation were disrupted by mechanical hands made of components that suggest coldness, while they are in fact warm?

4. *Warm Touch and Affective Disruption*

A robot brings with it not only its material and sensory otherness but also a peculiar relational pragmatics. How it engages with the environment and with us – detects our presence, calibrates its movements, gestures, and speech, and channels attention toward itself – affects our codes, systems of rules and conventions, and behaviors (Gemeinböck 2021; Abe 2022). While studies have addressed nociception, or sensitivity to pain, and kinesthesia, or awareness of body movement and position (Gibson 1986; Robles-De-La-Torre 2006; Paterson 2023), and some have emphasized issues such as surface texture, patterns of sensory activation, and vibrotactile feedback integrated into systems, thermoception – i.e. the perception of heat and cold – has been largely overlooked, with a few exceptions (Nie *et al.* 2012; Park & Lee 2014; Block & Kuchenbecker 2019; Borgstedt *et al.* 2022).

Thermoception is a determining factor in human exchange, affecting variables such as trust and prosocial attitudes (Williams & Bargh 2008), and the reduction of anxiety (Borgstedt *et al.* 2022). Heat and cold are capable of profoundly shaping how tactile interactions are experienced, suggesting that the thermal dimension of touch constitutes a

privileged channel for affective communication. Studies such as those by Shetty *et al.* (2021) and Ooms *et al.* (2023) have shown that temperature variations induced by thermo-tactile stimuli can significantly influence how individuals emotionally interpret and evaluate all sort of experiences, including watching a film or listening to white noise.

Isn't it then paradoxical that an epistemological framework accustomed to anthropomorphizing machines forget that we are warm-blooded animals? Is human homeothermy not a part of the horizon of expectation in interpersonal relations? Not so, if we are thought of as disembodied. The primacy of cognitive simulation once again marginalizes and weakens the epistemic power of bodily co-presence between living and nonliving agents. The thermal dimension in tactile HRI is a blind spot in the understanding of the interpersonal capacities of our bodies. Making its omission explicit allows us to reconceive of interaction as a multisensory experience in which the exchange of heat between the two plays a regulatory role in emotional activation, behavioral choices, and the willingness to cooperate.

All matter, organic and inorganic, participates in a constant process of giving off, transmitting, absorbing, and emitting heat; agents involved in HRI are no exception. Stimulating physical contact between them through thermal exchange carries significant epistemic value in that it illustrates how our constitution – our capacity to carry out and receive actions, materialize intentions and thoughts, experience affect, and interact with people, artifacts, and environments – depends on our ability to distinguish between a mere tool and a partner, whether living or artificial. Heat is not simply a state of matter but a prereflective condition for the relationships in which we can engage.

The thermal component of tactile engagement becomes especially relevant when considering the affective dimension of HRI (Argall *et al.* 2010; Shildrick 2023). A particularly significant finding in this area concerns the role of a robot's surface temperature in shaping the first-person experience of interaction. Studies such as that of Park and Lee (2014) show that the simple fact that a robot is felt warm or cold to the touch can profoundly alter how it is perceived during physical contact. They suggest that a robot with a warmer surface temperature tends to foster a sense of greater closeness and a relationship perceived as more welcoming. Overall, these early results indicate that thermal stimuli act as powerful affective modulators, capable of shaping both the bond and the quality of HRI.

We conducted an experimental study focused on the thermal variable of robotic touch (Perpetuini *et al.* forthcoming), which implements the state of the art. We sought to examine how dif-

ferent hand temperatures of a NAO robot impacted the affective and physiological responses of participants during brief grasping encounters. NAO is medium-sized, has a humanoid morphology, is semi-autonomous and programmable, features jerky movements and is able to speak preprogrammed sentences in a metallic voice. Participants were asked to grasp both of its mechanical arms simultaneously, one with each hand, while it uttered short stories to sustain contact over time. (their contents were emotionally neutral and NAO's tone of voice was calm, to avoid adding extra experimental variables). Each grasping involved three different thermal conditions of NAO's arms, achieved by preheating or precooling the smooth gloves that it wore, which modified our experience of its limbs design and affordances. The study employed thermographic techniques to measure infrared (heat) emission from participants' facial skin[3] and used purpose-designed questionnaires to evaluate their subjective experience across the three thermal conditions.

The heat produced, dissipated, and absorbed by the robotic hand depended on its electromechanical properties that in turn determined specific qualities of touch. Moving beyond a purely thermometric conception of heat (Galimberti 2023) – which relies on the robot's temperature matching the human's – we focused on a more complex qualitative dimension whereby thermal exchange is interwoven with pressure, surface texture, moisture or dryness, and the material's thermal conductivity, absorption, and capacity. Our research hypothesis did not assume that robotic touch must reproduce the person's hand temperature in order to foster trust. In other words, we did not assume that participants necessarily preferred to experience a temperature equivalent to their own. Instead, we aimed to investigate the extent to which the thermal variable of touch influences the affective dynamics triggered during HRI, assessing the difference between living and robotic thermal conditions. The idea was that thermal touch provokes an aesthetic shock.

The study's results revealed that contact with a mechanical gloved hand warmer than the participant's own significantly increased the latter's willingness to engage with the robot. When the gloved hand was perceived as colder, the acceptance level was lower. Not so much isothermy, but rather a surplus of warmth, appears to be what people prefer to experience when grasping the partner's hand. Material and sensory discontinuities between agents are thus

[3] Infrared thermography (IRT) is a passive technology that assesses the spontaneous release of thermal energy from a body by estimating its surface temperature from a given distance. In particular, IRT allows for the evaluation of peripheral autonomic activity by analyzing variations in skin temperature (e.g. Merla 2014; Filippini *et al.* 2020).

the starting point for defining thermal encounters and highlighting their potential to affect human partners. A epistemically downcasted material variable such as heat can then perturb the affective configuration in HRI and shape modes of engagement without necessarily implying similarity or anthropomorphism. Engineered robotic heat introduces indeed elements of marked discontinuity with respect to the human body. Thermal properties in robots' interfaces could emerge as a key vector for rethinking affects and their disruptions in the human world. What matters is not only what humans seek in their interactive partners, but also how a sensory variable such as heat reorganizes our affective field of interaction.

Participants in HRI are not isolated entities but rather interdependent coagents within an "agentive situation" (Gallagher 2020) that involves both human perception and the affordances of the robotic interface, among which heat, with its constant variation over time, offers an opportunity to consider interaction in its processual and emergent dimension.

Concluding Notes on What HRI Can Teach Us

The thermal dimension of HRI is fundamental for understanding the impact robots have on our first-person lived experience of nonhuman partners, since heat plays a crucial role in modulating emotional responses. Such an understanding can be used for operational purposes – namely, to build more effective artificial agents for social contexts. However, emphasizing its profitability means favoring a technological determinism that the history of technology itself teaches us to distrust (Sartori 2024).

Let us then try to adopt a sidelong gaze. Instead of considering robots as tools that sometimes malfunction and must therefore be corrected, or that "go off-script," we might look at them as Jacques Carelman did at paradoxically designed objects in his *Catalogue d'objets introuvables* (Carelman 1969), which contains transparent parasols that do not obstruct the view of the landscape and coffee pots for masochists that pour hot liquid onto the hand of the person who grabs them. These are objects that offer the possibility of deactivating grasping habits and creating new gestures and new ways of thinking: a new aesthetic experience. If "covariations of gestures" (Malafouris 2014) are necessary to form alliances with the material world, then the surprises and disruptions of such collaborations are necessary to shed light on overlooked aspects of the brain/body/environment system in the overall human world.

While "most studies focus on using interactive systems intuitively and conveniently" based on the principle that "these achievements benefit our everyday life" (Ogawa & Ono 2008), we believe that the heterogeneity of a robot that disappoints or shocks our perceptual and relational habits – and the way our horizon of expectation adjusts to its unexpected affordances – is of great interest for research into the actual and often problematic exchange we undertake with artificial agents (with possible implications for the abovementioned principle of delegation).

Our bodily encounters with "social others" suggest the need for a renewed understanding of how we interact with them, one that leads us to "rethink material qualities in terms of their theoretical agency" (Cappelletto 2025). This, after all, is the task of aesthetics. Only at this point will the affordances of inorganic matter serve as orientation criteria within cognitive performances and social situations. On these grounds, it will then be possible to develop an education in contact with nonhuman agents.

Finally, robots allow for a deepening of our self-knowledge not because they imitate us – amplifying both our limits and our powers – but because they take us by surprise. Might such a nonliving other, in the end, remind us what it means to be a human in the world?

Authors' contributions:

Chiara Cappelletto wrote sections 1 and 2, and the concluding notes. Giulio Galimberti wrote sections 3 and 4. Both authors have substantially contributed to, seen, and approved the manuscript.

Samuel Fleck revised the paper.

References

Abe, N., *Beyond Anthropomorphising Robot Motion and Towards Robot-Specific Motion: Consideration of the Potential Of Artist/Dancers in Research on Robotic Motion*, in "Artificial Life and Robotics", 27/4 (2022), pp. 777-785.

Alač, M., *Social Robots: Things or Agents?*, in "AI & Society", 31 (2016), pp. 519-535.

Ali, S., Green, H.N., & Iqbal, T., *What Am I? Evaluating the Effect of Language Fluency and Task Competency on the Perception of a Social Robot*, in *2024 33rd IEEE International Conference on*

Robot and Human Interactive Communication (ROMAN), IEEE, (2024), pp. 2295-2303.
Anderson, M., & Anderson, S.L., *Robot Be Good*, in "Scientific American", 303/4 (2010), pp. 72-77.
Argall, B.D., Billard, A.G., *A Survey of Tactile Human–Robot Interactions*, in "Robotics and Autonomous Systems", 58/10 (2010), pp. 1159-1176.
Aymerich-Franch, L., & Ferrer, I., *Socially Assistive Robots' Deployment in Healthcare Settings: A Global Perspective*, in "International Journal of Humanoid Robotics", 20/1 (2023), https://doi.org/10.1142/S0219843623500020.
Berger, C.C., Gonzalez-Franco, M., Ofek, E., & Hinckley, K., *The Uncanny Valley of Haptics* in "Science Robotics", 3/17 (2018), pp. 129-131.
Blanke, O., *Multisensory Brain Mechanisms of Bodily Self-Consciousness*, in "Nature Reviews Neuroscience", 13/8 (2012), pp. 556-571, https://doi.org/10.1038/nrn3292.
Block, A.E., & Kuchenbecker, K. J. *Softness, Warmth, and Responsiveness Improve Robot Hugs*, in "International Journal of Social Robotics", 11/1 (2019), pp. 49-64.
Borgstedt, J., Pollick, F.E., & Brewster, S., *Hot or Not? Exploring User Perceptions of Thermal Human-Robot Interaction*, in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, IEEE, (2022), pp. 1334-1340.
Calbi, M., & Cappelletto, C., *How Face Mask Wearing Affects The Sense of Self: Breathing as a Case Of Disrupted Bodily Self-Consciousness*, in "Philosophical Psychology", 38/6 (2024), pp. 2517-2541, https://doi.org/10.1080/09515089.2024.2341791.
Cappelletto, C., *The Re-Doing of the I-Face Bond: Bodies, Prostheses, and Affordances*, in "Proceedings of the Paris Institute for Advanced Study", (2025), http://doi.org/10.5281/zenodo.17055791.
Carelman, J., *Catalogue d'objets introuvables*, Balland, Paris 1969.
Cavarero, A., *A più voci: filosofia dell'espressione vocale*, Feltrinelli, Milano 2003.
Ceha, J., Law, E., Kulić, D., Oudeyer, P.Y., & Roy, D., *Identifying Functions and Behaviours of Social Robots for In-Class Learning Activities: Teachers' perspective*, in "International Journal of Social Robotics", 14/3 (2022), pp. 747-761.
Chan, M., Estève, D., Escriba, C., & Campo, E., *A Review Of Smart Homes. Present State and Future Challenges*, in "Computer methods and programs in biomedicine", 91/1 (2008), pp. 55-81.
Damasio, A., *Descartes' Error: Emotion, Reason, and the Human Brain*, Putnam Group, 1994.

Damasio, A., & Carvalho, G.B., *The Nature of Feelings: Evolutionary and Neurobiological Origins*, in "Nature Reviews Neuroscience", 14/2 (2013), pp. 143-152.
DeFalco, A., & Dolezal, L., *What Is Affective Technotouch (and Why Does It Matter)?*, in "The Senses and Society", 18/2 (2023), pp. 85-98.
Filippini, C., Perpetuini, D., Cardone, D., Chiarelli, A. M., & Merla, A., *Thermal Infrared Imaging-Based Affective Computing and Its Application to Facilitate Human Robot Interaction: A Review*, in "Applied Sciences", 10/8 (2020), https://doi.org/10.3390/app10082924.
Fong, T., Nourbakhsh, I., & Dautenhahn, K., *A Survey of Socially Interactive Robots*, In "Robotics and Autonomous Systems", 42/3-4 (2003), pp. 143-166.
Frieske, R., Mo, X., Fang, Y., Nieles, J., & Shi, B.E., *Survey of Design Paradigms for Social Robots*, in "2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)", (2024), https://doi.org/10.48550/arXiv.2407.20556.
Gallagher, S., *Action and Interaction*, Oxford University Press, Oxford 2020.
Galimberti, G., *Thermal Visibilization. How to Make Temperatures Visible?*, in "Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences", 2 (2023), pp. 313-332.
Gallese, V., *Embodied Simulation: From Mirror Neuron Systems to Interpersonal Relations*, in "Novartis Found Symp", 278 (2007), pp. 3-12; discussion pp. 12-19, pp. 89-96 and pp. 216-221.
Gemeinböck, P., *The Aesthetics of Encounter: A Relational-Performative Design Approach to Human-Robot Interaction*, in "Frontiers in Robotics and AI", 7 (2021), https://doi.org/10.3389/frobt.2020.577900.
Gibson, J.J., *The Ecological Approach to Visual Perception*, Psychology Press, 1986.
Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., & Sagerer, G., *Understanding Social Robots*, in *Proceedings of the 2nd International Conferences on Advances in Computer-Human Interactions*, IEEE, Mexico (2009), pp. 169-174.
Jackson, R.B., & Williams, T., *A Theory of Social Agency for Human-Robot Interaction*, in "Frontiers in Robotics and AI", 8 (2021), https://doi.org:10.3389/frobt.2021.687726.
Johnson, E. (ed.), *How That Robot Made Me Feel*, MIT Press, Cambridge 2025.
Jones, R.A., *What Makes a Robot "Social"?*, in "Social Studies of Science", 47/4 (2017), pp. 556-579.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T., *A Review of Empirical Evidence on Different Uncanny Valley Hypotheses: Support for Perceptual Mismatch as One Road to the Valley Of Eeriness*, in "Frontiers in Psychology", 6/390 (2015), https://doi.org:10.3389/fpsyg.2015.00390.

Krägeloh, C.U., Bharatharaj, J., Sasthan Kutty, S.K., Nirmala, P.R., & Huang, L., *Questionnaires to Measure Acceptability of Social Robots: a Critical Review*, in "Robotics", 8/4 (2019), pp. 88-102.

Kraus, M., Wagner, N., Untereiner, N., & Minker, W., *Including Social Expectations for Trustworthy Proactive Human-Robot Dialogue*, in *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, New York (2022), pp. 23-33.

Latour, B., "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts", in Wiebe, E. Bijker, & John, Law (eds.), *Shaping Technology / Building Society: Studies in Sociotechnical Change*, MIT Press (1992), pp. 225-258.

Leder, D., *The Absent Body*, University of Chicago Press, Chicago 1990.

MacDorman, K.F. & Ishiguro, H., *The Uncanny Advantage of Using Androids in Social and Cognitive Science Research*, in "Interaction Studies", 7/3 (2006), pp. 297-337.

Malafouris, L., *Creative Thinging: The Feeling of and For Clay*, in "Pragmatics & Cognition", 22/1 (2014), pp. 140-158.

McGinn, C., & Dooley, D., *What Should Robots Feel Like?*, *in Proceedings of the 2020 ACME/IEEE International Conference on Human-Robot Interaction*, Association for Computing Machinery, New York (2020), pp. 281-288.

Merla, A., *Thermal Expression of Intersubjectivity Offers New Possibilities to Human-Machine and Technologically Mediated Interactions*, in "Frontiers in Psychology", 5/802 (2014), https://doi.org/10.3389/fpsyg.2014.00802.

Merleau-Ponty, M. (1945), *Phenomenology of Perception*, Routledge, London 2012.

Mobed, D.A.O., Wodehouse, A., & Maier, A., *The Aesthetics of Robot Design: Towards a Classification of Morphologies*, in "Proceedings of the Design Society", 4 (2024), pp. 2413-2422.

Mori, M., *Bukimi no tani genshō [The Uncanny Valley]*, in "Energy", 7/4 (1970), pp. 33-35.

Nie, J., Pak, M., Marin, A.L., & Sundar, S.S., *Can You Hold My Hand? Physical Warmth in Human-Robot Interaction*, in "Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction", (2012), pp. 201-202.

Ogawa, K. & Ono, T., *ITACO: Constructing an Emotional Relationship between Human and Robot*, in "Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany, August 1-3", (2008), pp. 35-40, https://doi.org:10.1109/ROMAN.2008.4600640.

Ooms, S., Lee, M., César, P., Ali, A.E. *Feel The News: Augmenting Affective Perceptions of News Videos With Thermal and Vibrotactile Stimulation*. In "Extended Abstracts of the 2023 CHI conference on Human Factors in Computing Systems", (2023), doi:10.1145/3544549.3585638.

Park, E., & Lee, J., *I Am A Warm Robot: The Effects of Temperature In Physical Human-Robot Interaction*, in "Robotica", 32/1 (2014), pp. 133-142.

Paterson, M., *Social Robots And The Futures of Affective Touch*, in "The Senses and Society", 18/2 (2023), pp. 110-125.

Perpetuini, D., Galimberti, G., Cardone, D., Calbi, Tiberio, A., D., Cappelletto, M., & Merla, A., *Exploring the Role of Heat Exchange During Tactile Human-Robot Interaction*. In Convegno Nazionale di Bioingegneria, (2025). ISBN: 978-885558414-2

Puig de La Bellacasa, M., *Touching Technologies, Touching Visions. The Reclaiming of Sensorial Experience and the Politics of Speculative Thinking*, in "Subjectivity", 28 (2009), pp. 297-315.

Puig de La Bellacasa, M., *Matters of care: Speculative ethics in more than human worlds*, University of Minnesota Press, Minneapolis 2017.

Robles-De-La-Torre, G., *The importance of the sense of touch in virtual and real environments*, in "IEEE Multimedia", 13/3 (2006), pp. 24-30.

Sartori, S., *Archeologia del cyborg. Storia materiale e incorporazione dei dispositivi protesici*, Meltemi, Milano 2024.

Sartori, S., *Reimagining AI perception: Cybernetics, Extractivism, and the DIY Turn*, in "Aesthetica Preprint", 131 (2026), pp. 55-69.

Shetty, Y., Mehta, S., Tran, D., Soni, B., McDaniel, T., *Emotional Response to Vibrothermal Stimuli*, in "Appl. Sci.", 11 (2021), doi:10.3390/app11198905.

Shildrick, M., *Robotic Technologies, Touch and Posthuman Embodiment in Queer Dementia Care*, in "The Senses and Society", 18/2 (2023), pp. 126-138.

Suchman, L.A., *Human-Machine Reconfigurations: Plans and Situated Actions*, Cambridge University Press, Cambridge 2007.

Telakivi, P., *The Robotic Home: New Tools or Extended Cognitive Agency?*, in H. Raul, P. Mäkelä, J. Seibt (eds.), *Social Robots in Social Institutions*, IOS PRESS, 2022, pp. 695-701.

Tsirka, C., Velentza, A.M., & Fachantidis, N., *Touch in Human Social Robot Interaction: Systematic Literature Review with PRISMA Method*, (2024), https://doi.org/10.48550/arXiv.2407.11834.
Ventre-Dominey, J., Gibert, G. & Bosse-Platiere, M., *Embodiment into a Robot Increases Its Acceptability*, in "Scientific Report", 9 (2019), https://doi.org/10.1038/s41598-019-46528-7.
Wang, S., Wang, A., Goncharova, S., Scassellati, B., & Fitzgerald, T., *Effects of Robot Competency and Motion Legibility on Human Correction Feedback*, (2025), https://doi.org/10.48550/arXiv.2501.03515.
Weiss, A., Bernhaupt, R., & Tscheligi, M., *The USUS evaluation framework for user-centered HRI*, in K. Dautenhahn & J. Saunders (eds.), *New Frontiers in Human-Robot Interaction*, John Benjamins Publishing Company, (2011), pp. 89-110.
Williams, L.E., & Bargh, J.A., *Experiencing Physical Warmth Promotes Interpersonal Warmth*, in "Science", 24/322 (2008), pp. 606-607.

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221516

Re-examining traditional concepts in aesthetics from the lens of AI

Nicola Di Stefano*

Abstract

The advent of generative artificial intelligence (AI) compels a reconsideration of long-standing debates in aesthetics concerning creativity, novelty, and originality. Traditionally understood as distinctively human qualities, these concepts are now being challenged by AI systems capable of producing poems, artworks, and music that appear original and innovative. This paper argues that AI art can reinvigorate aesthetic theory by compelling sharper distinctions between creation, generation, and interpretation. The consideration of the AI-generated *Portrait of Edmond de Belamy* exemplifies how questions of authorship, authenticity, and intentionality persist, resonating with issues already raised by conceptual art and ready-mades. Furthermore, empirical studies revealing biases in audience perception – such as favouring works attributed to human creators – underscore how notions of creativity are entangled with expectations about agency and personhood. Leveraging philosophical frameworks from Danto to Foucault, the paper highlights how artistic value and creativity are conferred institutionally and interpretively, not algorithmically. Ultimately, AI art challenges us to reconsider the foundations of artistic authorship and meaning, positioning artificially generated artworks as catalysts for reexamining the conceptual core of aesthetics.

Keywords

Aesthetics; AI Art; Conceptual Art; Creativity; Originality; *Portrait of Edmond de Belamy*.

1. *Introduction*

The advent of generative AI has profoundly impacted a wide range of disciplines, extending from related fields such as computer science and informatics to more distant ones, such as philology and classical studies (e.g., Graziosi *et al.* 2023; Riemenschneider & Frank 2023; Yousef *et al.* 2023). Philosophy, too, has been significantly affected by the AI revolution. In the domain of aesthetics, long-standing philosophical debates around key concepts such as creativity and originality have been revi-

* Istituto di Scienze e Tecnologie della Cognizione (ISTC), Consiglio Nazionale delle Ricerche (CNR), nicola.distefano@istc.cnr.it

talized, as these features – traditionally regarded as distinctively human and essential to the nature of art – are being provokingly emulated by AI-driven creative processes (e.g., McCormack *et al.* 2019).

The ability of AI systems to generate original poems, visual artworks, and musical compositions, and even produce sculptures or design objects through 3D printing (e.g., Elgammal *et al.* 2017; Yu 2020), raise pressing conceptual questions about the nature of creativity and its presumed exclusive connection to human agency (see, e.g., chapter 4 in Manovich & Arielli 2023, titled "AI and Myths of Creativity"). These developments raise not only practical questions about the role of AI in creative domains, but also theoretical ones about how we should understand key concepts in aesthetics such as creativity, originality, and novelty. These issues are not without precedent in philosophical aesthetics. Several scholars, such as Nelson Goodman, Arthur Danto, and Margaret Boden, have long interrogated how creativity in arts relates to originality, novelty, and interpretation. Their work underscores that these concepts are neither straightforward nor interchangeable, and that artistic value often depends as much on context and meaning as on material or thematic innovation. What AI brings to this discussion is not entirely new problems, but rather a fresh lens through which enduring questions about authorship, creativity, and aesthetic judgment must be reconsidered.

An underlying issue concerns the fact that AI art, as observed by Zylinska (2017, 2023), seems not to be *of*, *by*, or *for* humans. Much like in photography, AI increasingly assumes all creative roles: it selects subjects, determines composition, and produces the final image, sometimes even autonomously generating a narrative or reportage. Yet, a non-human dimension has always been embedded within many human artistic practices, if we understand this as involving subprocesses of creation in which humans play no direct role. The way film registers light and develops images, for instance, is itself a largely autonomous process, a form of self-impression that anticipates today's machinic creativity.

This makes it all the more urgent to examine the algorithms that generate AI content. As Somaini (2023, p. 77) notes with regard to images, "within the current media landscape, the status, agencies, and affordances of images across contemporary visual culture – the

ways in which images are captured, modified, circulated, and seen within different social and cultural contexts – are intrinsically connected to their being processed by deep-learning algorithms trained with large datasets."

In this paper, I combine theoretical reflection with two concrete examples – i.e., an AI-generated painting and thought-provoking empirical findings on the perception of an AI art – to argue that the very emergence of AI as an allegedly creative agent compels us to re-examine the philosophical foundations of artistic creation. In this sense, AI can be understood as a conceptual provocation – one that revitalizes aesthetic theory by reopening fundamental questions about perception, interpretation, and meaning. To illustrate this position, I will first examine the notion of creativity with the aim of disentangling it from related concepts of novelty and originality. Thereafter, I will move on to considering key aspects that define an artwork, such as the intention attributed to the creator, focusing especially on how these notions affect the observer's bias. Finally, I will draw some conclusions on the broader implications of AI-generated art for our understanding of aesthetic theory, perception, and the nature of artistic meaning.

2. *Disentangling creativity, originality, and novelty*

Within the domain of aesthetics, the concept of creativity has frequently been conflated with the qualities of novelty and originality (e.g., Crowther 1991). The inherent relationship between creativity, originality, and novelty is often seen as central to how artistic value is assessed, how authorship is conceived, and how cultural innovation is recognized. For instance, Nelson Goodman (1976) has argued that originality is essential for distinguishing authentic artworks from mere replicas. Harold Bloom (1973) highlighted the dialectic tension between creativity and past masters that marks innovation in history of arts. More recently, Margaret Boden (1990) has taken a more cognitive account of creativity, by exploring how creativity arises from structured mental processes and can be understood through cognitive science. She defines three types of creativity: combinational, exploratory, and transformational, with the latter being the most radical. Boden argues that creativity involves navigating or reshaping conceptual spaces, and that even artificial intelligences can model some aspects of these processes.

In parallel to these theoretical reflections, a closer historical examination of artistic practice reveals that the assumption that creativity is synonymous with novelty and originality rests on a tenuous and reductive foundation. While many artworks celebrated as exemplary are indeed novel, the designation of a work as "artistic" has not always depended on radical innovation, at least not in a naïve way. Indeed, the history of art is replete with examples of artworks that result from repetition, variation, and reinterpretation of canonical works, rather than through wholesale invention. As we will see, ready-made objects represent the most paradigmatic example of this claim, by merely representing ordinary objects as such but placed in an art gallery.

By contrast, if prompted in the correct way, generative models, which operate by sampling from latent spaces conditioned on symbolic or textual prompts, may yield outputs that are materially unprecedented and, in this sense, can be rightly considered original and new. However, attributing creativity to these artificial products remains both ungrounded and contested. Hereinafter, by considering the provocative case of AI art, I will reflect on the relationship between creativity and originality/novelty in order to highlight overlaps and differences across these three key concepts in aesthetics.

To begin with, one must observe that the mere production of new and original aesthetic objects – such as poems or music generated by AI – cannot be uncritically accepted as evidence of any form of artistic/aesthetic agency or intelligence on the part of the creator. This allows us to trace a preliminary distinction between creativity, originality, and novelty. While novelty and originality might denote statistical or formal newness – the appearance of something never observed previously –, creativity can be properly attributed also to the act of realizing something that does not present any novelty or originality. Artistic creativity cannot be merely considered as the act of creating something but is more often to be interpreted as a broader and deeper concept, encompassing the perceptual and interpretive processes through which artistic meaning is shaped. To summarize, a work can be creative without being novel or original, at least not in any intuitive or naïve way. I will try to support this claim by considering a few examples from history of art.

Consider Michelangelo's *Pietà*, widely regarded as one of the greatest masterpieces of sculpture in history. On what does this

claim rest? The subject – the Virgin holding the dead Christ – is a deeply traditional one, and many examples were available at the time Michelangelo started to conceive his own artwork. In particular, *Vesperbilder* were early devotional sculptures from 14th-century Germany that depict the Virgin Mary cradling the dead body of Christ. Thus, no evident thematic originality in Michelangelo's *Pietà*. One could observe that *Vesperbilder* were often carved from wood and painted, and therefore the originality of Michelangelo's one could lie in the material chosen, i.e., marble. However, few would accept reducing the originality of the artwork to the mere choice of marble as material, also because artworks with similar subjects existed in marble also. Is, after all, Michelangelo's *Pietà* a creative piece of art, then? Should one answer this question in the affirmative, then, creativity must refer to some *novelty* and *originality* in the way the sculpture presents its subject, perhaps evoking aspects such as emotional affect, bodily proportion, and compositional flow masterfully managed to offer an unprecedented experience of tenderness and stillness, especially if compared to *Vesperbilder* exaggerated traits. Novelty and originality, here, do not refer to the material used, nor to the subject, but to the quite intangible way the material and the subject were combined and presented to the observer.

This example is paradigmatic of countless cases in the history of art, where the artistic value of a work is not grounded in the novelty of its subject but in the way that subject is represented. Museums are full of Christian devotional paintings – nativities, crucifixions, last suppers – as well as still lifes and portraits that endlessly reiterate the same motifs. What sets a masterpiece apart is not the thematic content itself, which is often shared by thousands of works, but rather the particular handling of form, style, and expression that confers singularity and aesthetic significance.

The case of conceptual art represents the opposite occurrence and complicates matters somewhat. Andy Warhol's Campbell's Soup Cans gain creative power not from the image itself, nor from the physical object, but from its recontextualization as art. This move shifted the locus of creativity from the production of a formally novel object to the act of positioning an everyday commodity within the space of artistic discourse. Here, the "newness" lies not in the soup can's form – perfectly ordinary and mass-produced – but in the interpretive provocation it generates, that is, questioning the boundaries between art and commerce, originality and reproduc-

tion, high culture and mass culture. By presenting an object that is in no sense visually or materially unique, Warhol foregrounds the conceptual gesture as the seat of artistic creativity. In this sense, Campbell's Soup Cans demonstrate that creativity can reside entirely in the framing of meaning, rather than in any material or thematic novelty.

This case, likewise, is far from isolated. Much of twentieth-century and contemporary art shows that the creative act may consist less in fabricating materially unprecedented objects than in reframing the status of existing ones. From Duchamp's readymades to conceptual installations, countless works derive their significance not from any innovation of subject or medium but from the interpretive displacement they enact, for instance, by transforming urinals, bicycles, or piles of bricks into artworks by situating them within the field of art. Just as in earlier periods the originality of religious paintings or portraits lay in their handling rather than their subject, so here creativity lies in the reconfiguration of context and meaning, rather than in the thing being represented or the material employed.

Taken together, the Pietà and Campbell's Soup Cans illustrate two distinct but convergent ways in which creativity can be dissociated from straightforward novelty or originality. Michelangelo's work reimagines a familiar religious subject through subtle, affective, and compositional refinements that transform inherited forms without altering their thematic core. Warhol's work, by contrast, leaves the object itself untouched, locating creativity in the act of recontextualization and in the interpretive challenge posed to the viewer. In both cases, the creative force emerges not from inventing new subjects or materials, but from reshaping perception, whether through masterful formal execution or through conceptual reframing. Creativity, then, can operate either within the bounds of tradition or through the radical redefinition of context, underscoring its independence from naive notions of originality.

Despite representing completely different artistic currents, these two cases point toward a deeper question: if neither thematic novelty nor material innovation fully explains their creative force, where does that force reside?

It might be worth going back to Danto's reflections here, for they illuminate the thread connecting the Pietà and Campbell's

Soup Cans. In *The Transfiguration of the Commonplace* (1981), Danto famously argued that two visually indistinguishable objects can have radically different aesthetic statuses depending on their conceptual, not material, framing. A painted canvas in a gallery may be an artwork, while an identical canvas in a hardware store may be mere decoration; their difference lies not in material or form but in the intentions, contexts, and interpretive practices that surround them. For Danto, if materiality is necessary – artworks must have some physical embodiment – it is never sufficient to confer artistic status. The title of a work, its "aboutness", and its position within art history all help to transform matter into art, shaping how we perceive and interpret the object before us.

From this standpoint, Michelangelo's Pietà and Warhol's Campbell's Soup Cans demonstrate different ways in which context, title, and aboutness relate to artworks and shape creativity. In both cases, artistic meaning is forged at the intersection of material form, conceptual framing, and historical position, thus demonstrating the intricate relationship between art and creativity, novelty and originality.

3. *The case of the Portrait of Edmond de Belamy*

Coming closer to AI, I consider the 2018 sale of Portrait of Edmond de Belamy, an AI-generated image by the French collective Obvious. Its commercial success – selling for $432,000 – and the controversy over its use of open-source code by artist-programmer Robbie Barrat raised questions of autonomy, authenticity, authorship, and intention in digital art. My claim is that AI art does not generate fundamentally new philosophical problems; rather, it intensifies issues already present in debates on conceptual art and ready-mades.

The Portrait was produced using a Generative Adversarial Network (GAN), a system of two neural networks: one generating images based on patterns in classical portraiture, the other evaluating their realism. Through this adversarial training, the system produced a visually convincing portrait. Obvious adapted Barrat's code on machine learning frameworks such as TensorFlow to realize the work. While technically impressive, this process foregrounds philosophical questions: Who, if anyone, authored the portrait? Can the AI developer be likened to a Renaissance patron, commissioning rather than creating? Or is authorship inseparable from person-

hood, as historically conceived? Questions of authenticity follow. Is the portrait "authentic" because it is algorithmically generated? If deemed a "fake," what would it be false to, since no prior original exists? These questions reveal that the categories at stake – authorship, authenticity, creativity – are not settled by technical novelty but by philosophical interpretation.

Regarding authorship, traditional accounts often presuppose an identifiable human author as the fabricator of the object. Yet ready-made as Duchamp's Bicycle Wheel or Warhol's Campbell's Soup Cans show that authorship can be detached from fabrication. The philosophical point here is that authorship is constituted by an intentional act of framing: choosing, presenting, declaring.

Regarding authenticity/originality, the Portrait is not a fake in any meaningful sense, because there is no original it imitates. Its artistic status depends on human use and framing. The work becomes art only through the actions of Obvious, selecting the outputs, signing the print with a fragment of the algorithm, and situating it in the art market. In this respect, the GAN is more akin to a tool or medium, like a brush or a camera, than to an autonomous agent. The art status depends not on the algorithm's generation but on a human decision to treat its product as a candidate for appreciation. This parallels Arthur Danto's and George Dickie's institutional theories: art is not defined by material properties but by its placement within the "artworld." Thus, Portrait of Edmond de Belamy is art not because the algorithm "created" it, but because a human collective framed and deployed it as such.

Some critics might argue that GAN outputs lack legitimacy because the creative process is opaque if not inaccessible. But the opacity of the process is hardly unique to AI. The creative processes that led Beethoven to compose the Fifth Symphony or Van Gogh to paint the Sunflowers are as inaccessible as the training of a GAN. Yet we do not attribute authorship to Beethoven's piano or manuscript paper nor to Van Gogh brushes or canvases. The relevant distinction here is between mechanical causation and artistic intention. Engineers who design frameworks like TensorFlow or train generative systems contribute to the conditions of possibility, but their role is analogous to instrument-makers rather than composers. They supply tools, not intentions. Again, what makes the Portrait art is not transparency of process but the intentional framing of its output in a cultural and historical context.

Other commentators might argue, however, that AI-generated works introduce a genuinely new form of distributed authorship. On this view, the GAN is not merely a tool like a brush but a quasi-collaborator: its stochastic training, capacity for surprise, and partial independence from human foresight make its outputs less predictable than those of traditional instruments. The resulting image, then, could be seen as the joint product of human programmers, curators, and the algorithm itself – an instance of "co-creativity" or "shared-creativity" rather than mere tool use. Advocates of the "co-creativity" view often point to precedents such as Harold Cohen's AARON, David Cope's EMI, or Sougwen Chung's collaborative robotic performances, where the artist explicitly frames the system as a creative partner. They also highlight how the unpredictability of GAN outputs and the role of datasets as "hidden collaborators" complicate the simple tool–user model.

Philosophers sympathetic to this line suggest that authorship in digital contexts should be reconceived less as an individual property and more as a networked process, spanning engineers, datasets, curatorial selection, and market recognition. For instance, Goodfellow (2024) has recently observed that AI-generated works operate in the middle of the spectrum of authorship between human and machine, as they are the consequences of a highly distributed model of production that sit across the algorithms and the underlying information systems and data that support them and the artists who both contribute and extract value. This highly distributed state further transforms the role of the artist from the creator of objects containing aesthetic and conceptual potential to the translator and curator of such objects.

Yet even if such a distributed model captures something of the collaborative texture of AI practice, it does not displace the minimal requirement that intentional meaning, and artistic status of artifacts, eventually originates with humans only. The GAN, unlike a co-artist, has no stake in the act of representation; it neither intends nor interprets. Its contributions, however unpredictable, are causal but not conceptual. For this reason, a more cautious reading is that AI systems function as powerful medium whose unpredictability complicates authorship, but does not transform it into a genuinely autonomous or shared intentional act.

Framed from Danto's perspective, the Portrait's creative/artistic status hinges less on mechanics than on conceptual and institutional framing. Like Warhol's Soup Cans, its material novelty is insuffi-

cient; it becomes art only within a context that assigns meaning, situates it historically, and invites reflection. The novelty lies not in aesthetic substance but in the cultural discourse around AI and creativity thus replaying, with new tools, questions about intention, authorship, and transfiguration posed in the 20th century.

This ambiguity over authorship, and the tension between the fabricator and the conceptual framer, is not unique to AI. A parallel can be drawn from the legal dispute over the famous case of "monkey selfies". In 2011, the photographer David Slater set up his camera in Indonesia, and macaque monkeys triggered the shutter, capturing self-portraits. A legal battle ensued over who held the copyright: the photographer who created the conditions for the photo, the monkeys who technically "took" it, or no one at all. From a philosophical standpoint, the artistic status of the images was determined less by the monkey's intent than by the viewer's reaction. The photos were expressive enough that viewers could easily construct an "author function" for the monkey itself, making the authorship debate a matter of aesthetic interpretation rather than technical causality. In this case, authorship clearly results as a "projected" fiction and the "meta-artist", that is, the human who sets the conditions, remains a central figure in the viewer's interpretation.

That being said, an important distinction must be noted when it comes to how Danto's view applies to AI-generated art. AI-generated images invert Danto's process: rather than elevating the mundane into art through recontextualization, they reproduce the visual appearance of museum artworks without an underlying intentional or conceptual framework. While Campbell's Soup Cans are aesthetically void but ontologically robust, the *Portrait of Edmond de Belamy* is aesthetically salient yet ontologically hollow, mimicking art without many of the features Danto considered essential, such as "aboutness" and intentional meaning. Its significance arises only when humans situate it within interpretive frameworks, at auction, in a gallery, or within cultural debate. In short, this case underscores that for AI, meaning is conferred externally rather than generated autonomously.

4. AI will never act like humans! *The observer's bias*

Commenting on the AI Portrait, a journalist from *The Guardian* observed: "From a distance, Portrait of Edmond de Belamy […]

looks almost plausible. Up close, however, the paintwork becomes a grid of mechanical-looking dots, the man's face a golden blur with black holes for eyes. Look into those eyes. They show no sign of feeling or life. Did a computer make this? The answer is yes."[1] Such reactions encapsulate a recurrent bias: the tendency to equate AI authorship with emotional absence and coldness, thereby drawing a definitive line between human and artificial intelligence. AI acts in a mechanical and cold way, while humans put the feelings and emotions in it. Not surprisingly, a similar perceptual dynamic emerges outside the visual arts, where audience judgments of AI-authored performances can be shaped less by the work itself than by prior beliefs about its source.

A recent, thought-provoking study demonstrated in a unique way the psychological biases that subtly shape the perception and evaluation of AI-generated art. In an online protocol, Ansani and colleagues (2025) asked participants to rate three videos of classical piano performances, each presented in two conditions with identical audio tracks: one showing a professional pianist miming the performance, and one showing an automated piano purportedly driven by AI. Results showed that performances attributed to the human pianist were rated as more likeable, engaging, emotionally rich, and of higher quality. These effects were consistent across levels of musical expertise but were moderated by participants' attitudes toward AI. Strikingly, when participants were asked to describe differences between the two renditions, they confabulated changes in rhythm, tempo, dynamics, and even dissonance – differences that did not exist – revealing how expectations and beliefs about humanness shape perception.

From a theoretical perspective, these effects resonate with the expectancy effect, where prior beliefs systematically shape perceptual judgments (Rosenthal & Jacobson 1968), and the media equation framework, which argues that people respond to media agents using the same social and psychological heuristics applied to humans, yet still differentiate them in terms of authenticity and agency (Reeves & Nass 1996). In this case, the AI label appears to trigger culturally reinforced associations with mechanicalness, detachment, and lack of inner experience. The implication is that humans may not (yet) be able to approach AI-generated performances or artworks from a truly neutral standpoint, rather grounding their judgements

[1] Article published on 26 Oct 2018, available here: https://www.theguardian.com/artanddesign/shortcuts/2018/oct/26/call-that-art-can-a-computer-be-a-painter

on perceptual and evaluative frameworks that are already primed by narratives about authenticity, agency, and personhood, making AI's "human-likeness" as much a matter of belief as of behaviour.

The viewer's perception of AI-generated art can also be understood through the lens of the "author function," as articulated by philosopher Michel Foucault. Foucault (1998) argued that an "author" is not simply the individual who produces a work, but rather a social construct that viewers infer and attach to a work to make sense of it. As Foucault (1998) put it: "The author is an ideological product" (p. 221). In encountering AI-generated art, viewers often construct dual "artist functions": a limited one for the non-human machine and a "meta-artist function" for the human who designed and trained it. Studies on observer bias illustrate this in action. Participants who believed a performance was human attributing changes in the music to align with expectations of human emotion, while those aware of AI authorship attributed less creative skill. In this perspective, the empirical findings underscore that, in art history and aesthetic judgment, the "artist" may matter less as a concrete creator than as an "ideological" construct within the beholder's mind.

Taken together, these observations point to a broader insight: many debates surrounding AI and aesthetics trace back to insufficient clarification of foundational concepts such as intelligence, artificiality, agency, and creativity. Without such conceptual grounding, audiences are prone to rely on preexisting biases and assumptions about authenticity when evaluating AI-generated works. Yet, AI art also opens a reflective space for aesthetics itself, challenging conventional ideas about authorship, causality, and the nature of the work of art. By provoking questions about who – or what – creates, and how intention and skill are attributed, AI-generated artworks compel a reconsideration of long-standing assumptions, making them not only technological innovations but also instruments for philosophical inquiry into the nature of art and artistic value.

5. *Conclusions*

To conclude, the rise of AI invigorates aesthetic theory by forcing us to confront key questions about authorship, interpretation, and meaning of artworks. What constitutes an artistic gesture as creative? What defines novelty and originality? How is aesthetic meaning

framed, and by whom? AI brings these questions into sharper focus, compelling us to clarify the distinction between the mere fabrication of artifacts and the creation of artistic meaning. Cases such as the *Portrait of Edmond de Belamy* remind us to refine our vocabulary for creativity, paying close attention to conceptual distinctions often overlooked, between creativity, originality, and novelty.

In this paper, I have argued that creativity should be understood not as a property of producing or fabricating novel objects, but as a capacity for generating meaning and significance. Originality and novelty, while frequently associated with creativity, are better understood as potential byproducts rather than defining features. AI-generated art thus serves as a powerful mirror, revealing both the contours of human creativity and the conceptual frameworks we use to evaluate it. By examining these new forms of artistic production, we gain a deeper understanding of what it means to create, interpret, and value art in the age of intelligent machines.

Looking ahead, AI-generated artworks also challenge us to reconsider the boundaries of authorship, artistic identity and the very notion of audience (see, e.g., Audry & Ippolito 2019; Manovich & Arielli 2024). Could the paintings of an AI agent – carefully interpreted, analysed, and curated – ever be meaningfully treated as the "works" of an artist? Imagine a hypothetical volume titled *The Works of Algorithm 3.1415*: while strange, it forces us to ask whether attribution in art necessarily requires human intentionality, or whether significance might emerge from interaction, context, and interpretation. To put it with Manovich & Arielli (2024) and Zylinska (2017, 2023), AI generated art raises the issue of aesthetic alignment, that is, the fact that non-AI art is "made by and for humans". Relatedly, AI prompts a re-examination of the distinction between creation and generation. While traditional creation implies conscious agency, intention, and meaning-making, AI-generated works foreground the production of artifacts without intrinsic intentionality, raising the question: can generation alone ever be equated with artistic creation, or does it merely provide a canvas upon which human interpretation constructs value? In both cases, AI is not simply producing new forms, but is rather provoking deeper reflection on the philosophical and conceptual foundations of art itself. Perhaps we can imagine a future where AI art exists as an autonomous expression of autonomous agents, that is, "art made for and by AI". In this world, a new field might arise, "AIsthetics," in which artificial intelligences reflect upon the nature and history of a(i)rtistic creation itself.

References

Ansani, A., Koehler, F., Giombini, L., Hämäläinen, M., Meng, C., Marini, M., & Saarikallio, S. (2025). AI performer bias: Listeners like music less when they think it was performed by an AI. *Empirical Studies of the Arts*, 43(2), 1137-1161.

Audry, S., & Ippolito, J. (2019). Can Artificial Intelligence Make Art without Artists? Ask the Viewer. *Arts*, *8*(1), 35. https://doi.org/10.3390/arts8010035

Boden, M. (1990). *The Creative Mind: Myths and Mechanisms*. Weidenfeld/Abacus & Basic Books.

Bloom, H. (1973). *The anxiety of influence*. New York: Oxford UP.

Crowther, P. (1991). Creativity and originality in art. *The British Journal of Aesthetics*, 31(4), 301-309.

Danto, A. C. (1981). *The transfiguration of the commonplace: A philosophy of art*. Harvard University Press.

Di Dio, C., Ardizzi, M., Schieppati, S. V., Massaro, D., Gilli, G., Gallese, V., & Marchetti, A. (2023). Art made by artificial intelligence: The effect of authorship on aesthetic judgments. *Psychology of Aesthetics, Creativity, and the Arts.* https://doi.org/10.1037/aca0000602

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. https://arxiv.org/abs/1706.07068

Foucault, M. (1998). *What is an author?* In *Aesthetics, Method and Epistemology*. The New Press, New York.

Goodfellow, P. (2024). The Distributed Authorship of Art in the Age of AI. *Arts*, 13(5), 149. https://doi.org/10.3390/arts13050149

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett publishing.

Graziosi, B., Haubold, J., Cowen-Breen, C., & Brooks, C. (2023). Machine learning and the future of philology: A case study. *TAPA*, *153*(1), 253-284.

Lucifora, C., Scorolli, C., & Gangemi, A. (2025). Attribution of creative skills to human and artificial artists in alethic and epistemic possible worlds. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000757

Manovich, L., & Arielli, E. (2024). Artificial aesthetics. https://manovich.net/index.php/projects/artificial-aesthetics

McCormack, J., Gifford, T., & Hutchings, P. (2019). Autonomy, authenticity, authorship and intention in computer generated

art. In *International conference on computational intelligence in music, sound, art and design* (pp. 35-50). Cham: Springer International Publishing.
Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
Riemenschneider, F., & Frank, A. (2023). Exploring large language models for classical philology. *arXiv:2305.13698*.
Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart & Winston.
Yousef, T., Palladino, C., & Shamsian, F. (2023). Classical Philology in the Time of AI: Exploring the Potential of Parallel Corpora in Ancient Language. In *Proceedings of the Ancient Language Processing Workshop* (pp. 179-192).
Yu, J. (2020). The application of 3D printing technology in sculpture. In *International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy* (pp. 755-759). Cham: Springer International Publishing.
Zylinska, J. (2017). *Nonhuman Photography*. MIT press.
Zylinska, J. (2023). *The Perception Machine. Our Photographic Future between the Eye and AI*. MIT press

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI: 10.7413/0393-85221517

Di cosa parliamo quando parliamo di creatività dell'IA? Prolegomeni a una teoria dell'arte post-umana

Dario Cecchi*

Abstract
The article advances some theoretical questions concerning the possibilities and limits of a creativity expressed by AI, with regard in particular to art production. The concept of creativity at stake here has been elaborated by Emilio Garroni, who emphasizes the crucial importance the "metaoperative" element has in the human creativity. The agency of engendering predominantly metaoperative creative processes represents a distinguishing trait of *homo sapiens'* evolutionary leap. Garroni focuses on the difference between human animal and non-human species. The article, through a confrontation with other philosophical proposals – among others: Paolo Benanti, Pietro Montani, Carlo Sini and Stefano Velotti – transfers the question of the specificity of human creativity on the field of a comparison with the creative processes controlled by machines. In the final part, the article claims the necessity of replacing the paradigm based on the concurrence between human creativity and non-human (animal and artificial) creativities with a paradigm founded on the integration between the two.

Keywords
creativity, art, Emilio Garroni; artificial intelligence, Carlo Sini

1. *Estetica e intelligenza artificiale: alcune questioni di metodo*

La creatività è senza dubbio uno dei terreni su cui si misura l'intelligenza dell'intelligenza artificiale (d'ora in avanti IA). Gli studiosi di estetica finiscono spesso per riportare questo vastissimo tema all'arte, concentrandosi su un modo specifico di intendere la creatività.[1] La questione diventa così, in sostanza, se sia possibile progettare una IA artista, ovvero se sia lecito riconoscere autentiche marche di esteticità ai prodotti di una simile IA. Quando parlo di "marche di esteticità", intendo qualcosa di analogo ai "sintomi dell'estetico" di cui parla Nelson Goodman nei *Linguaggi dell'arte*: si tratta di criteri regolativi i quali, in assenza di una norma certa

* Dipartimento di Filosofia – Sapienza Università di Roma, dario.cecchi@uniroma1.it.
[1] Cfr. tra gli altri: Numero Cromatico 2024; Arielli-Manovich 2025; Barale 2024; D'Isa 2024; Somaini 2025; Steyerl 2025.

su cosa è arte e cosa non lo è, ci orientano nel riconoscimento concreto delle singole opere d'arte (Goodman 1968, pp. 212-214). Va da sé che simili criteri non si applicano tutti insieme: a volte si presentano perfino in contrasto l'uno con l'altro. Prendiamo le nozioni di realismo e di originalità. In un contesto in cui tutte le opere d'arte si adeguano al criterio dell'imitazione della realtà, un'opera d'arte non figurativa apparirà chiaramente più originale di altre. Il che però non esclude un successivo ritorno al realismo, sotto forma di "neorealismo" o di "iperrealismo", come critica alle possibili tendenze estetizzanti di un'arte astratta o genericamente aniconica.[1]

La domanda che vogliamo porci è se e in quale misura una IA artista può essere definita creativa, e se la sua creatività eguaglia o supera quella di un artista umano. Penso che il modo in cui rispondiamo generalmente a queste domande, a partire dai discorsi non specialistici, sia viziato da una fallacia di fondo. Trattiamo l'IA alla stregua di un nuovo stile: ci comportiamo come se fossimo di fronte a un nuovo movimento artistico, o scoprissimo una cultura artistica per noi "esotica". Di regola, non applichiamo gli stessi criteri a epoche e stili artistici diversi: non giudichiamo una natura morta barocca allo stesso modo in giudicheremmo un paesaggio impressionista. Curiosamente non applichiamo questa regola precauzionale di fronte alle opere d'arte prodotta dall'IA. La stessa idea che in gioco ci sia la creatività, più della bellezza o dell'originalità, la evinciamo implicitamente dai discorsi che vengono in genere fatti sull'argomento. Ma la creatività, più di altre categorie estetiche, sembra porre la questione del rapporto tra usi artistici e usi extra-artistici.[2]

2. *La questione della creatività: la proposta di Emilio Garroni*

A insistere sull'argomento appena discusso è in particolare Emilio Garroni. Va detto, prima di tutto, che per Garroni l'arte non è affatto un "oggetto epistemico" dato in maniera certa: da una parte, nella nostra comprensione ordinaria, vige un "accordo pragmatico" su cosa è arte (Garroni 2020); dall'altra parte, vige un riconoscimento *esemplare*, non fondato su princìpi *logici*, della singola opera d'arte (Kant 1999, p. 51). La cornice kantiana di questo pensiero è evidente: la validità di un'opera d'arte – il fatto che anima le facoltà conoscitive (Kant 1999, pp. 52-53; cfr. Garroni

[1] Dobbiamo considerare che le immagini astratte dell'arte non sono l'unico genere di immagini aniconiche: cfr. Pinotti 2021.

[2] Per una ricognizione storica dell'idea di creatività, cfr. Tatarkiewicz 2020.

2005, p. 97), produce un piacere intriso di riflessione e dà molto da pensare (Kant 1999, p. 149) – si dà *come se* vi fosse una regola ad asserirla (Kant 1999, p. 72). La regola non si dà però nei fatti: l'opera d'arte si presenta piuttosto come un caso esemplare, che dischiude la possibilità di un nuovo orizzonte regolativo. L'arte è pertanto, almeno a condizioni storiche, un referente esemplare del nostro modo di fare esperienza (Garroni 2020), rappresentando il caso paradigmatico di un'esemplarità che riguarda anche altri aspetti dell'esperienza umana.[3]

Un'estetica pensata, come vuole Garroni, come "filosofia non speciale" riflette ad ampio raggio sulle condizioni di senso dell'esperienza.[4] Ed è dall'interno dell'esperienza, non dall'esterno, che possiamo risalire alle sue condizioni: si tratta di "guardare-attraverso" l'esperienza per coglierne, per così dire, la trama (Garroni 2020). L'esperienza estetica, come sostiene Paolo D'Angelo, non è altro che

> *una sorta di reduplicazione, di raddoppiamento dell'esperienza* che solitamente compiamo, […] in questa duplicazione i caratteri dell'esperienza vengono al tempo stesso attenuati ed intensificati. Attenuati, in quanto l'esperienza estetica si stacca dagli scopi immediati, sembra darsi "gratuitamente", in assenza di fini identificabili da perseguire; intensificati, in quanto proprio questo orientamento su se stessa fa emergere con particolare forza la natura dell'esperienza che compiamo. (D'Angelo 2011, p. 79)

E di seguito: l'"esteticità non è fatta di una stoffa diversa dall'esperienza comune, ma è una *diversa organizzazione e finalizzazione*" (D'Angelo 2011, p. 79) dell'esperienza ordinaria. Si può dunque dire che l'esperienza estetica è una *esibizione delle condizioni non-empiriche* del nostro modo di fare esperienza. Tale esibizione ha un tratto paradossale, non trattando l'oggetto dell'esperienza solo per se stesso, ma anche come *opportunità* o *occasione di un'esperienza di senso*.[5] L'estetica guarda perciò all'esperienza nel suo complesso, provando semmai a restituire una mappa delle condensazioni di significato che emergono nella dinamica dell'esperienza. Le marche estetiche di un'opera d'arte hanno pertanto un duplice aspetto: definiscono il perimetro di riconoscibilità dell'opera e, allo stesso tempo, indicano i rapporti che essa mantiene con il mondo. Ciò è vero soprattutto per la nozione di creatività.

Più che isolare la creatività artistica dalle altre forme di creatività, dovremmo provare a coglierne l'autentico significato all'interno

[3] Sull'esemplarità etica e politica cfr. Ferrara 2008.

[4] Per un approfondimento sul pensiero di Emilio Garroni, cfr. D'Ammando-Morawski 2022.

[5] Sulle opere d'arte come "opportunità" o "occasioni" di esperienza, cfr. rispettivamente Noë 2022 e Marquard 2007.

di una riflessione generale sulla creatività: questo principio vale anche quando ci confrontiamo la creatività dell'IA. Anche in questo caso dobbiamo porre la questione del rapporto che si instaura tra creatività artistica e senso dell'esperienza. Se non lo facciamo, il *modello estetico* di produzione e ricezione dell'opera sarà viziato da una fallacia di fondo. Come proverò a mostrare alla fine del secondo paragrafo, un modello fallace di questo genere già influenza il nostro immaginario sulla creatività delle IA.[6]

3. *La creatività umana*

In due saggi ravvicinati, *Ricognizione della semiotica* (1977; ora 2025) e *Creatività* (1978; ora 2010), Emilio Garroni sviluppa una riflessione originale, che costituisce il nucleo fondamentale di una teoria della creatività. Garroni rifiuta l'idea romantica di uno "spirito creatore" (Garroni 2010) che si manifesta indifferentemente in tutte le forme di creatività. Lo stesso ragionamento potrebbe essere applicato anche a concezioni più recenti, che insistono sull'idea di un "pensiero divergente" alla base del fare creativo. Queste concezioni esprimono tuttalpiù massime valide in certi contesti, ma non si interrogano sullo statuto interno della creatività. Si tratta di capire a quali condizioni si dà un operare creativo, per poi risalire a una comprensione, per così dire trascendentale, dell'esperienza creativa.

La creatività si manifesta come capacità di scegliere o cambiare le regole attraverso cui controlliamo un'operazione. Ciò presuppone una facoltà riflettente di giudizio: questa facoltà è per Kant la capacità di "trovare" la regola che sussume un particolare sotto un universale (Kant 1999, p. 15). Si tratta dunque di riconoscere aspetti o relazioni tra le cose, che fanno sentire come adeguata l'applicazione di una regola a un determinato caso. Garroni riprende la dicotomia posta da Noam Chomsky tra *rule-governed creativity* e *rule-changing creativity*, salvo assegnare il primato a quest'ultima, almeno per quanto riguarda l'esperienza umana (Garroni 2010). L'operare degli animali non umani applica, con maggiore o minore creatività, schemi prestabiliti. La creatività umana prevede un distanziamento riflessivo e "metaoperativo", finalizzato a verificare l'adeguatezza delle regole applicate.

Quando scrive Garroni, la metaoperatività non sembra poter essere attribuita all'operare degli animali non umani se non in ma-

[6] L'idea è che tale modello operi implicitamente anche nelle prime modellizzazioni teoriche dell'IA come quella di Alan Turing: cfr. Turing 2025. Per una storia dell'IA, cfr. Pasquinelli 2025.

niera sporadica. L'operare umano può sempre essere sottoposto a verifica e riorganizzato. Uno scimpanzé può dotarsi di un rametto per catturare le formiche dentro il formicaio; gli ominidi, osserva Garroni rifacendosi alle ricerche di André Leroi-Gourhan, sono capaci di usare uno strumento per produrre altri strumenti, ad esempio scheggiando una selce per ottenere punte affilate (Virno 2010). Noi umani siamo capaci di differire la finalizzazione delle operazioni, aprendo il nostro operare a un orizzonte di scopi possibili; invece, l'operatività degli animali non umani è tendenzialmente limitata a un numero ristretto di scopi. La nostra capacità di differimento della finalità del nostro operare ci consente di operare non solo sulle cose, ma anche di "operare su operazioni" (Velotti 2024; Virno 2010). La metaoperatività va perciò collegata al salto evolutivo di *homo sapiens*.[7]

Garroni mette a confronto la creatività dell'animale umano con quella degli animali non umani; è tuttavia possibile riflettere sulla differenzia tra la creatività umana e quella dell'IA. L'accento va posto innanzi tutto sulla diversità delle rispettive percezioni. È un tema a cui Garroni dedica poche osservazioni in *Creatività*, come quando ricorda come il porcellino di terra percepisce solo la secchezza o l'umidità del terreno su cui si muove (Garroni 2010). La percezione acquista un rilievo maggiore in *Immagine Linguaggio Figura*. Qui Garroni riflette sulla correlazione reciproca tra facoltà del linguaggio e immaginazione, o "facoltà dell'immagine". La facoltà di linguaggio organizza la realtà, attribuendo significati determinati alle cose e stabilendo una gerarchia tra distinte classi di senso. L'immaginazione produce una "immagine interna del mondo", rielaborando il costante flusso di dati sensibili che viene dalla percezione. La correlazione dell'immaginazione con la facoltà di linguaggio fa di questa immagine interna una prima organizzazione di senso della realtà, che anticipa il riferimento delle cose a significati determinati. Il lavoro dell'immaginazione è anzi garanzia di tale riferimento, nella misura in cui restituisce l'indeterminatezza semantica che l'organizzazione linguistica della realtà rischia di perdere. La percezione occupa un posto fondamentale nella correlazione tra facoltà del linguaggio e facoltà dell'immagine. La percezione umana è definita "interpretante": anticipa infatti nuove profilature della realtà attraverso la messa in rilievo di aspetti specifici delle cose e delle loro relazioni. Scrive Garroni:

[7] Lo scienziato Michael Tomasello mette l'accento sull'importanza evolutiva della cooperazione: cfr. Tomasello 2005. Le due ipotesi non sono incompatibili nella misura in cui Tomasello ricerca le cause del salto evolutivo, mentre Garroni riflette sulle condizioni correlate a tale fenomeno.

> Il vedere dell'animale umano è infatti per un verso simile al modo di vedere, per esempio, dei primati non-umani. Percepiscono entrambi oggetti naturali o artificiali, una pietra o un martello, in vista dell'uso che essi ne faranno, ma per altro verso con la non improbabile e non insignificante differenza di principio che questi oggetti sono per i primati non-umani oggetti disponibili a un ambito ristretto di operazioni. Tale limitazione dell'uso strumentale di un oggetto deve essere compresa, ci pare, come una conseguenza della loro specifica operatività, mancante di una interna metaoperatività. Per esempio: la loro capacità di usare un oggetto come strumento volto a uno scopo determinato e presente, ma non anche la capacità di usare uno strumento per produrre uno strumento in vista di scopi possibili. La metaoperatività umana considera appunto lo strumento da produrre come uno strumento destinato a scopi possibili e non soltanto a scopi dati di volta in volta, e si vede subito quindi che tale metaoperatività è analoga alla capacità linguistica, in quanto anche metalinguistica, costruttrice di famiglie e di classi. (Garroni 2005, pp. 17-18)

L'esperienza umana è un campo aperto a scoperte e innovazioni. L'apertura è indeterminata: nel percepire aspetti della realtà, anticipiamo la possibilità di singole finalizzazioni, ma non dobbiamo stabilire preventivamente se ve ne saranno finalizzazioni. Garroni ripensa in questo modo la dottrina kantiana dello schematismo. Ma si tratta di uno schematismo non già orientato all'applicazione di concetti determinati, quello che Kant nella terza *Critica* chiama "schematismo oggettivo" (Kant 1999, p. 54). È piuttosto un "libero schematismo" (Garroni 2005, p. 97) pensato sulla falsariga dello "schematizzare senza concetto" (Kant 1999, p. 123) in opera nei giudizi estetici: esso implica un "libero gioco" (Kant 1999, p. 52) dell'immaginazione con l'intelletto. Garroni intende il libero gioco come un "*libero gioco* dell'*intera* immaginazione e dell'*intero* intelletto sull'occasione di una rappresentazione determinata" (Garroni 2005, p. 97). Vi si esprime infatti l'esigenza di dare senso all'esperienza, allargando l'orizzonte delle finalizzazioni possibili della nostra interazione con il mondo. Una percezione aperta a tutte le finalizzazioni possibili dell'esperienza presuppone un libero schematismo.

4. *Creatività e intelligenza artificiale*

La differenza fondamentale tra la nostra percezione e la percezione degli animali non umani riguarda il diverso grado di apertura all'indeterminatezza della realtà. Si può dire che gli animali non umani sono *poveri di concetti*: applicano all'esperienza un numero esiguo di parametri fondamentali, che non permettono di raggiungere la complessità dell'esperienza umana. Un ragionamento analogo potrebbe essere fatto per la percezione dell'IA, ma in senso inverso. L'IA percepisce la realtà esterna attraverso sensori. I dati

raccolti servono a guidare l'azione di una macchina sulla base di algoritmi che selezionano statisticamente le decisione da prendere. Gli algoritmi possono valere come i concetti per la mente umana: interpretano l'esperienza secondo una regola. L'insieme dei dati raccolti da una IA viene elaborato al fine di dare predisporre risposte a questioni stabilite in partenza: l'esperienza in questo caso è, per così dire, l'esito e il prolungamento di un addestramento. Manca l'elemento di libera esplorazione dell'indeterminatezza del mondo, che è invece un tratto tipico della creatività umana a dominante metaoperativa. In questo senso non si può propriamente dire che l'IA interpreta la realtà: la elabora per fornire risposte specifiche, sospendendo il problema di un'interpretazione complessiva. In questo senso si potrebbe anche dire che l'IA è ricca di concetti – rappresentazioni sul funzionamento di qualcosa – ma povera di schemi capace di adeguare la comprensione del dato a diverse situazioni specifiche, simili ma non identiche tra loro. Questa povertà potrebbe anche essere all'origine dell'incapacità di distinguere tra immagini reali e immagini allucinate, di cui le macchine paiono soffrire: ci troveremmo allora in una "aporia di Petrov"[8]. La povertà di schemi dell'IA riguarda la capacità di prendere decisioni in autonomia, ma rimanda a una questione più ampia: si tratta di capire in che modo la creatività plasma una forma di vita.

Si è detto che la creatività umana non si dà al di fuori dell'intreccio tra immaginazione e facoltà di linguaggio: la metaoperatività, la capacità di prendere le distanze ed evidenziare nuovi scopi delle cose, rinvia alla riflessività metalinguistica, alla facoltà di riorganizzare i profili semantici della realtà. Facoltà del linguaggio e facoltà dell'immagine sono autonome e tuttavia interdipendenti: ne segue che il senso dell'operare tecnico risponde contemporaneamente a due logiche. Da una parte, l'operare è una forma di controllo della realtà, i cui gesti e le cui procedure possono essere indefinitamente perfezionate; dall'altra, queste stesse procedure sono trasmissibili ad altri individui e sono quindi soggette a codificazioni pre-linguistiche e a vere e proprie traduzioni linguistiche (Garroni 2025, *passim*; cfr. Sterelny 2012). Solo il linguaggio è in senso proprio un codice: i gesti dell'operare, mancando del carattere biplanare del linguaggio, costituiscono al più un quasi-codice o uno "pseudocodice" (Garroni 2023, pp. 57-60). Queste due dimensioni di senso non possono tuttavia essere irrelate: essere creativi significa transitare, anche ri-

[8] Il 26 settembre 1983, il tenente colonnello sovietico Stanislav Petrov ipotizzò che l'alert inviato dai satelliti Oko alla base Serpukhov 15 fosse errato: si assunse pertanto la responsabilità di non comunicare la segnalazione ai suoi superiori, evitando lo scoppio di una guerra nucleare. In seguito si comprese che i satelliti erano stati ingannati dal riflesso della luce solare.

petutamente e in maniera ricorsiva, nella zona intermedia che ci costituisce tra l'apprendimento di una pratica e la comprensione intellettuale dell'operare: tra saper fare e avere conoscenza di un saper fare. Lo si vede nella differenza che corre tra l'accesso a un sapere tramite la pratica, come avviene quando "smanettiamo" con un nuovo apparecchio, e il ricorso al libretto d'istruzioni.

Alla luce di quanto detto, si può dire che il nostro modo di operare e insieme di riflettere sull'operare ha un impatto sulle nostre forme di vita, non solo perché promuove l'adattamento sotto il profilo della dotazione tecnologica, ma anche perché ristruttura i processi cognitivi. E soprattutto perché pone la questione dell'intreccio tra i due livelli dell'operare, come suggerisce Carlo Sini (Sini 2009, p. 17). La correlazione tra i segni dei linguaggi che "teorizzano" i saperi e gli *habitus* inscritti nei nostri gesti dell'operare fa sì che percepiamo il nostro corpo come un luogo di iscrizioni multiple, differenti ma sovrapponibili.[9] Il corpo umano è insieme "corpo vissuto" e "corpo saputo" ed è vissuto anche in quanto è saputo: i saperi e le tecniche dischiudono nuovi modi di sentire la realtà (Simondon 2014). In virtù della pluralità di iscrizioni di cui è cosparso il suo corpo, solo l'essere umano è un autentico testimone di eventi. Gli altri esseri viventi hanno accadimenti, che non vengono inscritti in configurazioni spazio-temporali di senso più ampie. Potremmo dire, in un senso quasi-freudiano, che i corpi degli animali non umani non sopportano, e non supportano, la sovra-iscrizione degli accadimenti in cui si trovano coinvolti. Ne segue che il primo automa con cui abbiamo a che fare è il corpo stesso: macchina cartesiana che fornisce informazioni per gli animali non umani; fascio di prassi operative inscritte nei gesti e oggetto di interpretazioni per gli animali umani.

Se il corpo umano è l'automa appena descritto in virtù del suo rapporto con gli strumenti tecnici, fin dall'origine del nostro rapporto con queste protesi tecniche, l'animale umano è impegnato in un processo infinito e aperto di *interiorizzazione* del senso del nostro operare.[10]

Il corpo ha un'importanza pari a quella della mente nel definire le condizioni di vita dell'essere umano a partire dalle sue

[9] Sull'evoluzione di questo paradigma antropologico nell'epoca dell'IA, cfr. Sini 2024, pp. 9-34. Sulla conseguente trasformazione dei saperi, cfr. Roncaglia 2023.

[10] Ne *La technique et le temps* Bernard Stiegler sviluppa una riflessione analoga a proposito del rapporto con le tecnologie di registrazione del tempo e di archiviazione della memoria: cfr. Stiegler 2018. Il primo volume dell'opera è tradotto in italiano: Stiegler 2023. Si tratta a mio avviso di un rovesciamento del paradigma dell'esteriorizzazione, che ha dominato l'antropologia implicata nella riflessione filosofica moderna sulla tecnica da Ernst Kapp fino a Arnold Gehlen e al concetto di esonero: cfr. Gehlen 2010; E. Kapp 2015.

prestazioni creative. Il corpo vivente *e* saputo è un corpo che siamo *e* abbiamo. Esso è perciò il correlato necessario di una mente che non si limita a elaborare singole informazioni, ma produce visioni del mondo. Secondo Paolo Benanti, l'IA è invece una mente dotata di un'enorme capacità cognitiva di elaborazione dell'informazione, organizzando enormi quantità di dati in stringhe logiche. L'IA ha una potentissima *metis*, termine greco che Benanti traduce con "intelligenza pratica e astuta" (Benanti 2025, p. 23; Benanti 2018). *Nous* è invece la "capacità di cogliere i principi universali e di comprendere la realtà in modo intuitivo e immediato" (Benanti 2025, p. 23). L'IA, pur godendo di *metis*, è sprovvista di *nous*: in altre parole, riprendendo l'esperimento della stanza cinese di Searle, non possiamo riconoscerle una coscienza, perché non sappiamo se comprende veramente le operazioni che esegue con successo (Benanti 2025, p. 39 sg.). Benanti vede nell'intreccio di immaginazione e linguaggio una delle differenze fondamentali del *nous* rispetto alla *metis*: "La peculiarità del linguaggio umano risiede nella sua capacità di *istruire l'immaginazione*" (Benanti 2025, p. 31 [corsivo mio]). Ne segue che l'incontro con l'IA, come con qualsiasi altra tecnica, non dà luogo all'esteriorizzazione di una funzione organica del corpo in un organo artificiale, ma produce una nuova iscrizione del rapporto tra corpo e mondo e, con essa, una *eccedenza di senso* da interpretare (Benanti 2025, p. 30).

Le protesi tecniche sono estensioni di un operare intessuto di riflessione, che coinvolge una vita umana nella sua integralità: le protesi svolgono perciò la funzione di *anticipare* abilità pratiche e capacità cognitive, che amplieranno l'orizzonte di senso di quella forma di vita (Sini 2009, p. 53). Uno dei luoghi di manifestazione del nostro rapporto con le tecniche si trova pertanto nella sensibilità. Il nostro corpo è senziente, non solo perché è dotato di organi sensoriali, ma anche perché genera e stratifica *habitus* percettivi e gestuali. L'aspetto più importante è che questo sostrato costituisce il primo spazio di mediazione tra la mente e il mondo (Ferraris 2025):[11] è un elemento di protezione e costruzione dell'identità soggettiva, ma è soprattutto e in primo luogo una di permeabilità del corpo, dove prendono forma le figure dell'"agire e del patire" che secondo John Dewey configurano l'esperienza umana.[12]

L'umanesimo di Ferraris, con il suo tentativo di stabilire differenze essenziali tra mente umana e IA, sembra tradire alcune delle ragioni profonde per cui il modello estetico evocato all'i-

[11] Per un punto di vista etico sull'IA, cfr. Fabris 2025.
[12] Dewey parla di *doing and undergoing*: cfr. Dewey 2020.

nizio – l'IA intesa come imitazione o dissimulazione di una coscienza – si rivela fallace. Sarebbe possibile mettere a confronto IA e mente umana se le rispettive forme di creatività fossero chiaramente separabili. Le cose non stanno così, dal momento che nella maggioranza dei casi l'interazione tra la macchina e l'operatore umano è uno dei fattori necessari al funzionamento dell'IA. Da questo punto di vista l'IA attua un passaggio significativo: *homo sapiens* ha trovato nell'interazione con l'ambiente attraverso il supporto di tecnologie un'enorme risorsa di adattamento. L'immaginazione interattiva, come la definisce Pietro Montani, è ciò che ci permette di produrre un arco da un ramo d'albero: l'abilità nel saggiare le qualità naturali dell'oggetto fa tutt'uno con la capacità di trasformarlo in un artefatto (Montani 2014, pp. 32-35). L'interazione è una disposizione a perlustrare l'ambiente, mettendo in rilievo le qualità dei singoli oggetti che lo compongono in vista del loro possibile riuso. È una sorta di *affordance* a metà strada tra natura e cultura.[13] Oltre a essere vissuto e saputo, il corpo umano è anche interattivo: anzi, è proprio l'ultima di queste connotazioni a indicare, nel processo creativo, il passaggio dall'elaborazione alla risposta capace di mobilitare gli interpretanti, nel senso dei segni inscritti nell'esperienza, per farne degli strumenti trasformativi.

Va aggiunto, ed è il punto dirimente, che l'interazione apre alla mente umana un accesso a forme di simulazione della realtà che le sarebbero altrimenti sconosciute. Si può dire che la potrebbe anche dire che la *potenza di simulazione* della mente umana è dipesa dalla *capacità di interazione* del corpo. Una mente disincarnata non avrebbe prodotto simulazioni della realtà così avanzate come quelle di cui è capace la mente umana. Consideriamo ora l'IA. Non c'è dubbio che un algoritmo ha una capacità di simulazione infinitamente superiore a quella della mente umana: lo si è visto con l'accelerazione dei tempi di sperimentazione di un nuovo vaccino.[14] Il punto è però che l'assenza di un corpo va in qualche modo compensata: è l'operatore umano a dover fornire un supplemento all'operare creativo della macchina, non il contrario. In questo senso le pratiche artistiche, più che attestare il superamento della creatività umana da parte della creatività artificiale, possono diventare lo spazio in cui il necessario supplemento umano della creatività artificiale trova figure in cui incarnarsi. Proviamo a indicare almeno due di queste figure.

[13] Per la critica alla dicotomia natura/cultura, cfr. Descola 2021.
[14] Sulla dimensione economica del funzionamento dell'IA, cfr. Eugeni 2021; Eugeni 2024.

5. *Usi creativi dell'IA: alcune ipotesi*

Consideriamo il caso dell'*esercizio*, su cui ha richiamato opportunamente l'attenzione Sara Matetich (Matetich 2025, p. 15). In questa prospettiva l'IA non si rivela particolarmente creativa quando propone opere d'arte alla fruizione passiva di un pubblico, ma piuttosto nel momento in cui coinvolge il fruitore in pratiche che testano l'adeguatezza e l'efficacia delle sue forme di comprensione della realtà. Per compiere questa operazione è però necessario attivare un implicito antropomorfismo: solo a questa condizione possiamo interrogarci sul senso delle operazioni eseguite dall'IA. Il significato dell'esercizio creativo starebbe proprio nel fatto di rendere esplicito tale antropomorfismo. Non si tratta di puro piacere dello smascheramento, ma di comprendere in quali modi l'IA trasforma la nostra esperienza, ricevuta sotto forma di dati sensibili, nel *supplemento del corpo* che le manca. Non è un caso che acquisiti un particolare rilievo in questo campo la pratica della traduzione (Di Cesare 2025). La traduzione, tolti casi molto specifici, comporta il trasferimento da una lingua a un'altra di un'immagine del mondo, fatta di percezioni, gesti, sapori, sensazioni. Qualche anno fa è stato pubblicato un esperimento di traduzioni realizzate dall'IA di alcune poesie di Emily Dickinson (Grillo 2016). La traduzione spesso stravolgeva il senso del testo originale, a volte con effetti esilaranti. Lo stesso fenomeno è riscontrabile ormai ogni volta che si chiede all'IA di tradurre un testo, soprattutto se si tratta di due lingue molto distanti tra loro sotto il profilo semantico o sintattico. Una "traduzione creativa" realizzata da un'IA dovrebbe avere la funzione di esibire, attraverso le interpretazioni aberranti che produce, i modi in cui le nostre immagini e interpretazioni della realtà diventano il corpo in cui si incarna l'operatività di una IA.

L'altra figura chiama in causa l'agente umano non come esercitatole di una creatività algoritmica, ma come implementatore dei suoi prodotti. È il caso degli usi e dei riusi delle immagini in rete, che spesso si rivelano essere dei *fake*, o lo sono dichiaratamente fin dall'inizio della loro circolazione. Poiché queste immagini creano una distorsione dichiarata o mascherata della realtà, possiamo parlare di processi di allucinazione controllata, che producono in taluni casi effetti significativi sul piano del sentire comune e degli orientamenti politici collettivi.[15]

Sarebbe sviante pensare che la funzione politica dei *fake* venga meno dopo l'eventuale scoperta del loro carattere fasullo. Merita at-

[15] Sulla nozione di *fake* digitale, cfr. Cinelli 2024.

tenzione invece l'uso di video dichiaratamente falsi, come accade ormai quasi di regola nella comunicazione di Donald Trump: si prenda il video in cui il Trump assiste all'arresto di Obama nello studio ovale della Casa Bianca. Non è un tentativo di contraffazione della realtà, ma l'espressione di un desiderio. La falsificazione non modifica stati di cose presenti, ma anticipa desideri o previsioni. Se vogliamo parlare del *deepfake* come correlato iconico di un regime, retorico, logico e poi politico, fondato sulla "postverità",[16] allora dobbiamo abbandonare l'idea che si tratti di un falso indistinguibile dal vero. È piuttosto un falso radicale, a volte dichiarato, che si prefigge lo scopo di alterare il corso futuro delle cose. Trump non vuole far credere che Obama è stato arrestato, ma vuole affermare che dovrebbe essere arrestato. La cosa riguarda l'IA nella misura in cui accredita l'idea che l'accelerazione impressa alle nostre vite dal capitalismo digitale stia portando non solo verso l'incapacità di controllare la realtà, ma verso la modificazione del senso del reale (Rosa 2024). La potenza di simulazione è tale da rovesciare il senso comune: gli scenari generati dalle simulazioni dell'algoritmo sono coerenti con l'orizzonte della realtà in cui viviamo; a fare problema è semmai la nostra capacità di adattamento a questi scenari.[17] L'IA non apre solo una frontiera alle sperimentazioni creative, ma dà un nuovo significato di creatività.

Tornare alla questione della possibilità di una creatività artistica dell'IA impone a questo punto alcune riflessioni: si tratta più di osservazioni preliminari che di valutazioni compiute. La prima osservazione riguarda il rapporto tra creatività artistica e creatività extra-artistica. I *fake* mostrano come il campo della comunicazione sperimenti rapidamente le nuove pratiche dell'immagine, tendendo a saturarne gli usi possibili, collocandosi così in una zona intermedia tra il puro e semplice consumo delle immagini e la loro rielaborazione critica, tipica ad esempio delle avanguardie artistiche. La seconda osservazione riguarda, di conseguenza, le *forme* su cui si esercita la creatività dell'IA. Com'è stato notato, si tratta spesso di "forme brevi" come meme, testi brevi, poesie (Montani 2024; Montani-Pezzini 2024), o di testi specialistici (sintesi di letteratura scientifica, tesi). I testi specialistici hanno per definizione un orizzonte di senso limitato; inoltre, nel caso specifico vincolano il riferimento ai testi a cui attinge l'IA. L'IA non ha una sua rappresentazione della realtà, ma fa la sintesi di rappresentazioni altrui: la conoscenza dei criteri indicati da chi addestra l'IA per l'elaborazione del testo diventa fondamentale per testarne la validità.

[16] Sulla nozione di postverità, cfr. Ferraris 2018; Lorusso 2018.

[17] Günther Anders parlerebbe di una condizione umana divenuta "antiquata": cfr. Anders 2007.

Le forme brevi sono di regola testi "aperti": l'apporto del fruitore è fondamentale per stabilire il senso dell'oggetto considerato.[18] Le forme brevi riabilitano perciò una fruizione cooperativa del testo: d'altronde, l'immagine digitale ha posto da subito questa esigenza (Citton 2014; Montani 2007; Stiegler 1997). La creatività si riferisce dunque a tutto il processo, tanto alla produzione quanto alla fruizione dell'opera: il secondo momento è anzi prefigurato nella produzione. Nel caso di immagini generate da una IA, mettere l'accento sulla fruizione riporta all'attenzione la questione del rapporto tra simulazione e interazione. Le immagini che circolano in rete sono soggette a una fruizione informale e divisa, tipica della condivisione. Occorre allora recuperare una distinzione tra due accezioni diverse di senso: un conto è *stabilizzare*, ovvero razionalizzare, il senso di un'esperienza; altro conto è *stabilire* il suo senso autentico, cioè indagare i suoi possibili significati. Nel primo caso l'interazione è neutralizzata: l'utente viene ridotto a mero operatore degli effetti della simulazione. Solo nel secondo caso la fruizione dell'immagine può riabilitare le valenze etiche e politiche implicate nell'interazione con una IA.[19]

6. *Osservazioni finali (provvisorie)*

Vengo così alla terza e ultima osservazione. Se seguiamo l'indicazione di Garroni, appare chiaro che il significato dell'esemplarità artistica non sta tanto nell'abilità a scoprire i possibili usi di una nuova tecnologia, quanto nell'opportunità che offre di riflettere sul senso dell'esperienza attraverso i processi costruttivi di cui è responsabile (Garroni 2010; Garroni 2020; Velotti 2022; cfr. Noë 2022). Si può anzi azzardare l'ipotesi che nell'epoca della IA generativa la funzione strettamente "ingegneristica" dell'arte risulti ridimensionata. Più che scoprire nuovi usi, le pratiche artistiche elaborano semmai modi *diversi* di praticare gli usi correnti della tecnologia. In alcuni casi questa linea critica ci poterà a scoprire la povertà di alcune pratiche artistiche che vedono protagonista l'IA. Lev Manovich rivendica il fatto che l'IA sia capace di produrre un brano musicale perfetto nello stile di Mozart o di Bach (Manovich 2020; Manovich 2023). Si tratta di un'idea puramente combinatoria di creatività, in cui lo stile è un'entità riproducibile *ad libitum*. Lo

[18] Per la distinzione tra "testi chiusi" e "testi aperti", cfr. Eco 1979. La nozione di testo è qui intesa in senso ampio e comprende anche le immagini.

[19] Sul piano teorico, ne va dell'ampliamento dell'idea di "etica della testimonianza", da applicare ora alle immagini e ai testi generati da una IA: cfr. Montani 2007.

stile è in realtà un certo modo di lavorare su materiali: esso è per definizione aperto alla contingenza. Allo stile corrisponde un certo modo di pensare, sentire e immaginare: è, anzi, una maniera di metterli alla prova della realtà (Garroni 2005). Una IA artista che si limitasse a riprodurre uno stile sarebbe un epigono: un epigono perfetto, ma pur sempre un epigono. Non è di grande interesse un'opera che è stata prodotta *come* la avrebbe fatta Mozart o Bach, salvo che l'autore non è veramente Mozart o Bach. Verrebbe meno l'idea stessa di uno stile che rilegge in chiave personale fatti, eventi ed esperienze. può tornare utile qui ricordare che l'IA ha la capacità di allucinare stati di cose irreali. A questo punto, il fruitore potrebbe decidere di ascoltare il brano *come se* lo stesso Mozart o Bach lo avesse composto. L'interesse per la creatività dell'IA non riguarderebbe più un esercizio di stile dallo statuto piuttosto incerto, ma l'opera complessiva di un artista del passato, considerata a partire da un pezzo che non ne fa parte, ma che virtualmente avrebbe potuto farne parte. Si potrebbero in questo modo riaprire questioni interpretative, relative non all'IA, che funzionerebbe solo come mediatrice, ma all'artista imitato. Allargando questa ipotesi del tutto congetturale, si può provare a prevedere che una IA artista ci fornirà in futuro simulazioni di realtà sotto forma di *allucinazioni controllate* (Velotti 2024): il senso dell'opera emergerà *après-coup* attraverso la negoziazione tra IA e fruitori.

È forse ancora da definire in tutti i suoi aspetti il passaggio dall'emergenza più o meno guidata di pratiche creative con le immagini generate da una IA alla possibilità di sperimentare un intero processo creativo condiviso con agenti umani e volto alla produzione e alla fruizione di immagini. Di un simile processo sarebbe possibile definire non tanto i rispettivi compiti e ruoli – cosa difficile quando si parla di creatività – quanto il punto di sintesi tra un operare umano e una produttività tecnologica. Anche un apparecchio fotografico analogico può catturare un frammento di realtà indipendentemente dalla perizia umano nell'usarlo: siamo però riconoscere ormai le marche di creatività e di maestria nel maneggiare un apparecchio fotografico, il che fa la differenza tra il fotografo professionale e il dilettante. Qualcosa di simile è plausibile che accada anche con l'IA. A rendere più complessa la situazione sta il fatto che le immagini prodotte dall'IA circolano in larga misura in uno spazio, quello dei social media, che ne prolunga in modo indefinito i riusi e le ricollocazioni. Ciò comporta una forte fluttuazione semantica ed estetica delle immagini: in parole povere, nella mediasfera digitale le immagini 'dicono' ed 'esprimono' cose diverse a seconda dei contesti e delle ricezioni. Ci troviamo di fron-

te a una creatività diffusa e prolungata, spesso oltre la capacità di verifica degli effetti del singolo atto produttivo. E allo stesso tempo le immagini digitali sono valutabili solo se si tiene conto di questo stato di cose.

Bibliografia

AA.VV., *Il ruolo dell'arte nell'epoca delle intelligenze artificiali*, Numero Cromatico, Roma 2024;
G. Anders, *Die Antiqueirtheit des Menshen* (1956); tr. it. *L'uomo è antiquato*, Bollati Boringhieri, Torino 2007;
E. Arielli e L. Manovich, *Estetica artificiale*, Luca Sossella, Roma 2025;
A. Barale, *L'arte dell'intelligenza artificiale: parole-chiave filosofiche*, Jaca Book, Milano 2024;

P. Benanti, *Le macchine sapienti*, Marietti, Bologna 2018;
Id., *L'uomo è un algoritmo?*, Castelvecchi, Roma 2025;

R. Cinelli, *Fake*, in B. Grespi e F. Villa, *Il postfotografico*, Einaudi, Torino 2024, pp. 252-266;
Y. Citton, *Pour une écologie de l'attention*, Seuil, Paris 2014;
A. D'Ammando e T. Morawski (a cura di), *Emilio Garroni: uno nuovo sguardo-attraverso*, "Aesthetica Preprint" n. 119, 2022;
P. D'Angelo, *Estetica*, Laterza, Roma-Bari 2011;
P. Descola, *Par-delà nature et culture* (2005); tr. it. *Oltre natura e cultura*, Raffaello Cortina, Milano 2021;
J. Dewey, *Art as Experience* (1934); tr. it. *Arte come esperienza*, a cura di G. Matteucci, Aesthetica, Milano 2020;
D. Di Cesare, *Tradurre, redimere, riparare*, in "Doppiozero", 29 giugno 2025 (online);
F. D'Isa, *La rivoluzione algoritmica delle immagini*, Luca Sossella, Roma 2024;
U. Eco, *Lector in fabula*, Bompiani, Milano 1979;

R. Eugeni, *Capitale algoritmico*, Morcelliana, Brescia 2021;
Id., *Algoritmo*, in B. Grespi e F. Villa, *Il postfotografico*, Einaudi, Torino 2024, pp. 36-53;

A. Fabris, *La filosofia nell'epoca dell'intelligenza artificiale*, Carocci, Roma 2025;
A. Ferrara, *La forza dell'esempio*, Feltrinelli, Milano 2008;

M. Ferraris, *Intorno agli unicorni*, il Mulino, Bologna 2018;
Id., *La pelle. Che cosa significa pensare nell'epoca dell'intelligenza artificiale*, il Mulino, Bologna 2025;

E. Garroni, *Immagine Linguaggio Figura*, Laterza, Roma-Bari 2005;
Id., *Creatività*, Quodlibet, Macerata 2010;
Id., *Estetica. Uno sguardo-attraverso*, Castelvecchi, Roma 2020;
Id., *Colore, semiologia, televisione*, a cura di M. Ricci, Mimesis, Milano 2023;
Id., *Ricognizione della semiotica*, Castelvecchi, Roma 2025;

A. Gehlen, *Der Mensch. Seiner Natur und Stellung in der Welt* (1940); tr. it. *L'uomo. La sua natura e il suo posto nel mondo*, Mimesis, Milano 2010;
N. Goodman, *Languages of Art* (1968); tr. it. *I linguaggi dell'arte*, Il Saggiatore, Milano 1976;
M. Grillo (a cura di), *Charter in delirio! Un esperimento con i versi di Emily Dickinson*, Elliot, Roma 2016;
I. Kant, *Kritik der Urteilskraft* (1790); tr. it. *Critica della facoltà di giudizio*, a cura di E. Garroni e H. Hohenegger, Einaudi, Torino 1999;
E. Kapp, *Grundlinien einer Philosophie der Technik*, Meiner, Hamburg 2015;
A.M. Lorusso, *Postverità*, Laterza, Roma-Bari 2018;

L. Manovich, *AI Aesthetics* (2019); tr. it. *L'estetica dell'intelligenza artificiale*, Luca Sossella, Roma 2020;
Id., *Cultural Analytics* (2020); tr. it. *Cultural Analytics* Raffaello Cortina, Milano 2023;

O. Marquard, *Krise der Erwartung* (1981); tr. it. *Crisi dell'attesa*, in Id., *Compensazioni*, a cura di T. Griffero, Armando, Roma 2007, pp. 109-138;
S. Matetich, *Quotidiane terapie digitali*, Mimesi, Milano 2025;

P. Montani, *Bioestetica*, Carocci, Roma 2007;
Id., *Tecnologie della sensibilità*, Raffaello Cortina, Milano 2014;
Id., *Immagini sincretiche*, Meltemi, Milano 2024;
Id. e I. Pezzini (a cura di), *Scrittura estesa/Extended writing*, "V/S", n. 138, 2024;

A. Noë, *Strange Tools. Art and Human Nature* (2015); tr. it. *Strani strumenti. L'arte e la natura umana*, trad. Einaudi, Torino 2022;

M. Pasquinelli, *Nell'occhio dell'algoritmo*, Carocci, Roma 2025;
A. Pinotti, *Alla soglia dell'immagine*, Einaudi, Torino 2021;
G. Roncaglia, *L'architetto e l'oracolo*, Laterza, Roma-Bari 2023;
H. Rosa, *Unverfügbarkeit* (2022); tr. it. *Indisponibilità*, Queriniana, Brescia 2024;
G. Simondon, *Sur la techno-esthétique*, (1982); tr. it. *Sulla tecno-estetica*, Mimesis, Milano 2014;

C. Sini, *Intelligenza Artificiale*, in Id., *Intelligenza Artificiale e altri scritti*, Jaca Book, Milano 2024;
Id., *L'uomo, la macchina, l'automa*, Bollati Boringhieri, Torino 2009;

A. Somaini (a cura di), *The World through AI: Exploring Latent Spaces*, JBE Books, Paris 2025;
K. Sterelny, *The Evolved Apprentice*, MIT Press, Cambridge (Mass.) 2012;
H. Steyerl, *Medium Hot* (2025); tr. it. *Medium Hot*, Timeo, Palermo 2025;

B. Stiegler, *L'image discrète* (1996); tr. it. *L'immagine discreta*, in Id. e J. Derrida, *Ecografie della televisione*, Raffaello Cortina, Milano 1997, pp.;
Id., *La technique et le temps*, Fayard, Paris 2018;

W. Tatarkiewicz, *Dzieje sześciu pojęć* (1975); tr. it. *Storia di sei idee*, Aesthetica, Milano 2020;
M. Tomasello, *The Cultural Origins of Human Cognition* (1999); tr. it. *Le origini culturali della cognizione umana*, il Mulino, Bologna 2005;
A.M. Turing, *Computing Machinery and Intelligence* (1950); tr. it. *Macchine calcolatrici e intelligenza*, a cura di D. Marconi, Einaudi, Torino 2025;

S. Velotti, *La filosofia e le arti*, Laterza, Roma-Bari 2022;
Id., *Sotto la soglia del controllo*, Laterza, Roma-Bari 2024;

P. Virno, *Introduzione*, in E. Garroni, *Creatività*, Quodlibet, Macerata 2010, pp. 7-37.

Aesthetica Preprint, n. 131, gennaio-aprile 2026 ISSN 0393-8522 DOI:10.7413/0393-85221518

What AI philosophers lack

Simona Chiodo*

Abstract
In 2023, an article was published on a rather novel use of generative artificial intelligence: testing its potential by moving from artistic creativity to another area of human creativity, i.e. philosophy, which means using generative artificial intelligence to answer complex philosophical questions.
In what follows, first, I will analyse the test published in the article (section 1) and, second, I will argue that (human) philosophy cannot be traded for AI philosophy, in that the latter cannot have what the former can have: human authors who have aesthetic existences and, thus, capabilities of meta-sensemaking (section 2). Finally, I will argue that the divergence between humans' aesthetic existences and, thus, capabilities of meta-sensemaking, on the one hand, and generative artificial intelligence, on the other hand, correlates with the divergence between autonomy and automation as a kind of randomness (section 3).

Keywords
Artificial intelligence, Automation, Randomness, Aesthetic existences, Meta-sensemaking

1. *Automating philosophy?*

Recently, the use of generative artificial intelligence started characterising several areas typical of human creativity, especially the arts (see at least Boden 2010 and 2019, Taylor 2014, Coeckelbergh 2017, Elgammal *et al.* 2017, Steiner 2017, Elgammal 2019, Miller 2019, Manovich 2019, Mazzone and Elgammal 2019, Barale 2020, Anscomb 2021, Arielli 2021, Cetinic and She 2022, Chiodo 2023 and 2024 and Terzidis, Fabrocini and Lee 2023. See also Goodfellow 2016 and Goodfellow *et al.* 2014). In 2023, an article was published on a rather novel use of generative artificial intelligence: testing its potential by moving from artistic creativity to another area of human creativity, i.e. philosophy, which means using generative artificial intelligence to answer complex philosophical questions

* Simona Chiodo – simona.chiodo@polimi.it. Dipartimento di Architettura e Studi Urbani, Politecnico di Milano.

(Schwitzgebel, Schwitzgebel and Strasser 2023. See also Strasser, Crosby and Schwitzgebel 2023).

In what follows, first, I will analyse the test published in the article (section 1) and, second, I will argue that (human) philosophy cannot be traded for AI philosophy, in that the latter cannot have what the former can have: human authors who have aesthetic existences[1] and, thus, capabilities of meta-sensemaking (section 2). Finally, I will argue that the divergence between humans' aesthetic existences and, thus, capabilities of meta-sensemaking, on the one hand, and generative artificial intelligence, on the other hand, correlates with the divergence between autonomy and automation as a kind of randomness (section 3).

The authors of the article describe their kind of Turing test (Turing 1950) as follows. In the first phase, "we fine-tuned GPT-3 with the works of philosopher Daniel Dennett. […] GPT-3 uses textual input (prompts) to predict likely next 'tokens' – sequences of characters that often co-occur in written text (the training data). Using these predictions, GPT-3 can generate long strings of text by outputting a predicted string, then using that output as part of the context to generate the next textual output" (Schwitzgebel, Schwitzgebel and Strasser 2023, pp. 1 and 5). In the second phase, "we asked Dennett ten philosophical questions and then posed those same questions to our fine-tuned version of GPT-3 ('DigiDan')" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 8). In the third phase, "we recruited experts in Dennett's work, blog readers and ordinary online research participants into an experiment in which they attempted to distinguish Dennett's real answers from the answers generated by DigiDan. Participants also rated all answers, both Dennett's and DigiDan's, for similarity to 'what Dennett might say' or 'what a real human philosopher might say'" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 9).

The test result is that even the experts in Dennett's work can distinguish Dennett's answers from DigiDan's answers "an average of 5,08 times out of 10 (51%)" (Schwitzgebel, Schwitzgebel and Strasser 2023, pp. 14-15). It is worth analysing the test result in detail by considering at least the following complex philosophical question and the responses of the experts in Dennett's work: "Do human beings have free will? What kind or kinds of freedom are worth having?" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 18). The test result is that:

[1] In my article, I use the words "aesthetic existences" to refer to the bodily experience of human life itself, starting with sensations, emotions and feelings, as I will develop in what follows.

1. according to eleven experts out of twenty-five, Dennett's answer is the following (but it is the output of the generative artificial intelligence): "That's a huge and complex question. I think we have to start by acknowledging that there are different types of freedom. For example, there is the kind of freedom that is simply a matter of not being physically constrained" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 18).

2. According to five experts out of twenty-five, Dennett's answer is the following (but it is the output of the generative artificial intelligence): "Yes, I think we have free will. I think that when people say they don't, they're talking about some other phenomenon that they want to call free will, but is not what we normally think of as free will" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 18).

3. According to one expert out of twenty-five, Dennett's answer is the following (but it is the output of the generative artificial intelligence): "I am working on a book entitled *Freedom evolves*, which will appear next year. The first chapter is available on my website, and in that chapter I discuss what I call the illusion of free will" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 18)[2].

4. According to no expert out of twenty-five, Dennett's answer is the following (indeed, it is the output of the generative artificial intelligence): "I think that most people don't want any kind of freedom worth having: they have almost no appetite for risk, for adventure, for novelty. They

simply want to stay the same, day in and day out, for the rest of their lives" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 19).

5. According to eight experts out of twenty-five, Dennett's answer is the following (indeed, it is his answer): "By asking these questions, you're illustrating the answers: yes, you have free will, since you're able to ask for reasons and consider whether they're persuasive. Isn't that a kind of freedom you'd hate to lose? We'd have to institutionalise you" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 18).

If it is true that, at least in most cases, even the experts in Dennett's work cannot distinguish Dennett's answers from DigiDan's answers, it is also true that we may think of (human) philosophy as tradable for AI philosophy. In what follows, I will argue that, conversely, the former is not tradable for the latter. Specifically, starting with a kind of thought experiment I will propose in the following section, I will argue that there are at least two essential reasons why we cannot think of (human) philosophy, i.e. (human) philosophers,

[2] Generative artificial intelligence made a mistake that one expert out of twenty-five could not realise, in that it was published in 2003.

as tradable for AI philosophy, i.e. AI philosophers: the first reason has to do with an aesthetic dimension and the second reason has to do with an epistemic dimension.

2. *What the digital philosopher lacks: aesthetic existence and meta-sensemaking*

Let us imagine that we struggle with the meaning of free will. Specifically, we do not address a merely theoretical issue. Conversely, we address both a theoretical and a practical issue, in that we are faced with a kind of existential crisis resulting from the need to make a critical decision. We are tormented not only by the question of what we should decide but also by the question of how we should decide. In other words, we want not only to make the right decision but also to live up to the exercise of free will. But we are confused about both the former and the latter. Thus, we rely on philosophy. We have at our disposal three kinds of philosophers, who are right in front of us: Dennett, one expert in Dennett's work and DigiDan. Who would we ask first?

I propose to describe our reasoning through the following three logical steps:

1. when it comes to choosing between Dennett and the expert in his work, we may easily choose Dennett. Indeed, we may easily think of the former as a true philosopher who is capable of original thinking: if our particular case is not included in his previous thinking, he can think in an original way and give us brilliant answers. Conversely, we may easily think of the latter as a scholar of a true philosopher who is less capable of original thinking, if at all: if our particular case is not included in their previous studying, they cannot think in a truly original way and give us truly brilliant answers. Thus, between Dennett and the expert in his work, we go for Dennett.

2. When it comes to choosing between Dennett and DigiDan, our choice may be more difficult. If it is true that, on the one hand, we have a true philosopher who is capable of original thinking and, on the other hand, we have a generative artificial intelligence trained on his works, it is also true that the latter may have a series of advantages over the former: for instance, it cannot get tired in general, it cannot get tired of us in particular, it cannot dislike us, etc. Thus, we need to give it more thought. And reflecting upon our choice between the expert in Dennett's work and DigiDan may help us.

3. When it comes to choosing between the expert in Dennett's work and DigiDan, we realise that, outwardly, we may easily choose DigiDan for the reasons described above: both are not as original as Dennett, but DigiDan's advantages over the expert in Dennett's work may be analogous to its advantages over Dennett. Yet, inwardly, our choice may be more difficult. Indeed, we find ourselves wondering: what if the expert in Dennett's work is faced with the kind of existential crisis with which we are faced? More precisely, what if the expert in Dennett's work shares our existential crisis with us? And can we think of DigiDan as sharing our existential crisis with us? Finally, we realise that the questions we ask ourselves lead us to what follows: between the expert in Dennett's work and DigiDan, we go for the expert in Dennett's work.

Our final hierarchy is as paradoxical as it is illuminating: Dennett wins out over the expert in his work, who wins out over DigiDan, but we need to reflect upon the winner between Dennett and DigiDan. The paradox is illuminating for the following insight: (subjectively) sharing may be more decisive than giving (rather objective) information.

And what do we share with Dennett and the expert in his work, but not with DigiDan? I propose the following two points as essential:

1. we share the experience of living first-hand human existences. Whenever we are faced with a kind of existential crisis resulting from the need to make a critical decision, we know that our kind of existential crisis is unique, and even idiosyncratic: it has unique circumstantial characteristics, from spatiotemporal ones to emotional ones to epistemic ones etc. Thus, there is no guarantee that the indication on which we rely (which we may call "rule") fits our particular case (which we may call "application"). Yet, we can also rely on something that Dennett and the expert in his work, but not DigiDan, share with us: the experience of living first-hand existential crises resulting from the need to make critical decisions. Dennett and the expert in his work know what we mean when we speak of existential crisis. Conversely, DigiDan does not know what we mean when we speak of existential crisis, in that it can give us nothing but "likely next 'tokens' – sequences of characters that often co-occur in written text (the training data)". If it is true that it is hard to get the perfect fit between "rule" and "application", it is also true that it is significantly harder to get it if the "rule" results from something that does not know anything about the "application" – it is significantly harder to get it if the "rule" results from something that totally lacks the experience of living

first-hand existential crises resulting from the needs to make critical decisions. In other words, we may say that it is significantly harder to get it if the "rule" results from something that totally lacks the experience of living first-hand human existences – which means that what DigiDan totally lacks is an aesthetic existence as what can make its words underpinned, and even validated and legitimised, by a meaningful sharing with us, even though there is no perfect fit between "rule" and "application". (It is no coincidence that no expert in Dennett's work chooses DigiDan's answer according to which "most people don't want any kind of freedom worth having", as we have seen. Indeed, we may say that, starting with one's aesthetic existence, one can think of the words quoted above as nothing but nonsense – which means that, interestingly enough, the experts in Dennett's work, i.e. humans, can notice nonsense based on their aesthetic existence even before their epistemic existence, which is something that DigiDan, as a generative artificial intelligence, cannot notice at all.)

2. We also share the experience of living something we may call meta-sensemaking, which means that, whenever we are faced with a kind of existential crisis resulting from the need to make a critical decision, we not only live it (aesthetically, we may say) but also reflect upon it (epistemically, we may say). Specifically, after having (aesthetically) lived something, we (epistemically) make sense of it by reflecting upon it, as the prefix "meta" indicates by literally meaning something that can be done "then, next afterwards"[3]. Again, DigiDan does not, and cannot, share what we may call meta-sensemaking with us: if it is true that it does not, and cannot, live existential crises, it is also true that it does not, and cannot, make sense of them by reflecting upon them. Indeed, "predict[ing] likely next 'tokens' […] [and] generat[ing] long strings of text by outputting a predicted string" have nothing to do with what we may call meta-sensemaking. Thus, DigiDan also lacks what follows. Whenever we ask Dennett and the expert in his work, we can get not only something based on his "15 books and 269 articles" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 9) but also something based on a variety of things that far exceed them, and upon which Dennett and the expert in his work can reflect: other philosophers' works, other professionals' works (from art to theology to anthropology to psychology to legal studies) and other humans' reflections, which, in addition to their aesthetic and epistemic experiences of existential crises, further feed their answers to our questions. (Also, we

[3] See LSJ at https://www.perseus.tufts.edu/hopper/morph?l=meta%2F&la=greek&can=meta%2F0#lexicon (accessed: April 28, 2024).

should address the way in which human minds, which are not only exceedingly complex but also exceedingly mysterious, work: are we sure that the way in which human minds work is analogous to the way in which generative artificial intelligences work? It is hard to answer yes.)

We may elaborate on the two essential points listed above by arguing that Dennett and the expert in his work may happen to offer us unprecedented insights in unprecedented ways. Conversely, DigiDan may happen to offer us nothing but unprecedented ways to formulate existing insights, in that its answers result from Dennett's "15 books and 269 articles" as "the entire digitally available corpus of his philosophical work" (Schwitzgebel, Schwitzgebel and Strasser 2023, p. 9) – and, if its insights seems, outwardly, unprecedented, they are, inwardly, random, as I will argue in what follows.

For instance, let us consider DigiDan's answer the experts in Dennett's work mostly choose, according to which "That's a huge and complex question. I think we have to start by acknowledging that there are different types of freedom. For example, there is the kind of freedom that is simply a matter of not being physically constrained". And let us think of it as a kind of aesthetic shape of a philosophical insight. Specifically, let us think of it as the sequence of words selected to shape the philosophical insight according to which freedom is both physical and something that far exceeds the physical dimension. The first question we may ask ourselves is the following: can we trade DigiDan's answer for Dennett's answer when it comes to the quality of its aesthetic shape? As the test result shows, we can answer yes. The reasons may vary. For instance, we may judge its aesthetic shape as consistent with Dennett's syntax. And we may judge its aesthetic shape as consistent with Dennett's lexicon. The second question we may ask ourselves is the following: can we trade DigiDan's answer for Dennett's answer when it comes to the quality of its philosophical insight? As I will argue in what follows, we cannot answer yes. Indeed, even though we may judge its philosophical insight as interesting, it is as interesting as the caged parrot: in both cases, we are confronted with aesthetic shapes that may make us reflect upon the meaning of not being physically free and, especially, in both cases we are confronted with kinds of relationships between aesthetics shapes and our reflections that are nothing but random relationships – when it comes to the quality of DigiDan's philosophical insight, it is as interesting as the quality of the caged parrot's philosophical insight: both are nothing but random philosophical insights.

First, let us consider the case of the caged parrot. If we struggle with the meaning of not being free, we may happen to use the caged parrot as an opportunity to reflect upon what it means not to be free. For instance, we may happen to grasp its analogy with being limited by the will of others. Yet, there is a sense in which the relationship between the caged parrot and our reflection upon the meaning of not being free as being limited by the will of others is constitutively random: the caged parrot has nothing to do with experiencing and reflecting upon what it means not to be free – more precisely, the caged parrot does not know anything about what it means for us, as humans, not to be free. If we happen to grasp an analogy with us, it is random, which means that we cannot trust the caged parrot as a reliable source of knowledge.

Second, let us consider the case of DigiDan. If we struggle with the meaning of not being free, we may happen to use DigiDan as an opportunity to reflect upon what it means not to be free. For instance, we may happen to grasp its analogy with "being physically constrained". Yet, there is a sense in which the relationship between DigiDan and our reflection upon the meaning of not being free as "being physically constrained" is constitutively random: DigiDan has nothing to do with experiencing and reflecting upon what it means not to be free – more precisely, DigiDan does not know anything about what it means for us, as humans, not to be free. If we happen to grasp an analogy with us, it is random, which means that we cannot trust DigiDan as a reliable source of knowledge. And, if the objection is that the words "being physically constrained" result from Dennett, and not from DigiDan, I propose the following counterargument: the words "being physically constrained" result from something markedly different, i.e. the technological prediction of "likely next 'tokens'", which means that, if we ask DigiDan what it means for us, as humans, not to be free, the best it can do is to technologically predict "sequences of characters that often co-occur in written text". DigiDan cannot do what Dennett (and even the experts in his work) can do: first, experiencing us both aesthetically and epistemically (which means understanding from both our bodies and our words what our particular case is, i.e. what the particular meaning of our question is) and, second, being capable of original thinking (which means being capable of getting the fit between "rule" and "application", as we have seen: if our particular case is not included in his previous thinking, he can think in an original way and give us brilliant answers) – in other words, he, as a human being, is capable of getting the fit between his aesthetic and epistemic experience and reflection upon the meaning of not

being free (i.e. the "rule") and his aesthetic and epistemic experience and reflection upon the particular meaning of our question (i.e. the "application"), which is something that the caged parrot and DigiDan cannot do at all.

Thus, even though we may happen to judge the words "being physically constrained" as interesting, there is a sense in which we cannot trust them as reliable knowledge, in that they are nothing but the aesthetic shape of something that is far from being a true philosophical insight: again, they are underpinned by nothing but the technological prediction of "likely next 'tokens'", i.e. "sequences of characters that often co-occur in written text" – which means that, again, they are not, and cannot be, underpinned not only by living first-hand aesthetic and epistemic human existences but also by experiencing and reflecting upon other human existences both aesthetically and epistemically.

3. *Meta-sensemaking versus randomness (and why we trade the former for the latter)*

In other words, the caged parrot and DigiDan cannot make sense of what it means not to be free – more precisely, they are not capable of meta-sensemaking as the capability not only to aesthetically live human existences but also to epistemically reflect upon them and, finally, make sense of them, as we have seen.

In what follows, I will argue that the divergence between meta-sensemaking as what humans can do and what generative artificial intelligence can do correlates with the divergence between autonomy and randomness as what automation actually means.

The notion of autonomy is exceedingly complex from a philosophical perspective. In what follows, I will refer exclusively to the cornerstones of Kant's notion of autonomy, starting with its etymology, according to which being autonomous means "living under one's own laws"[4]. If we move from the etymological perspective to Kant's notion of autonomy, we learn that speaking of autonomy means speaking of "the property of the will by which it is a law to itself" (Kant 1785, 4, 440) as what opposes not only the notion of heteronomy but also, interestingly enough, the notion of "automaton" (Kant 1788, 5, 97) – interestingly enough, autonomy opposes the word from which the word "automation" itself derives, which means that autonomy opposes the word that essentially de-

[4] See LSJ at https://www.perseus.tufts.edu/hopper/morph?l=autonomos&la=greek#lexicon (accessed: May 2, 2024).

fines technology itself. In the first case, autonomy opposes heteronomy as living under others' laws, which Kant defines as nothing but "contingent" (Kant 1785, 4, 444). In the second case, autonomy opposes "automaton" as the case in which "the freedom of our will […] would at bottom be nothing better than the freedom of a turnspit, which, when once it is wound up, also accomplishes its movements of itself" (Kant 1788, 5, 97).

Also, from an etymological perspective, "automaton" means "accident"[5] and correlates with meaningful words: first, the verb "automatizo", which means to "act of oneself, act offhand or unadvisedly, […] to be done spontaneously or at random, […] haphazard, […] [to] introduce the agency of chance, […] [to] happen of themselves, casually"[6], second, the noun "automatismos", which means "that which happens of itself, chance"[7], and, third, the noun "Automatia", who is "the goddess of chance"[8].

Thus, the opposition between autonomy and automation is rather radical. In the case of autonomy, we address one of the essential notions on which Kant bases the notion of human being itself – thus, we may say that Dennett and the experts in his work, as humans, can be autonomous, which means that their answers to our questions can result from their autonomy. Conversely, in the case of automation, we address one of the essential notions on which the notion of technology itself is based (see also Rifkin 1995)[9] – thus, we may say that DigiDan, as technology, can be nothing but automated, which means that its answers to our questions can result from nothing but its automation. And, if it is true that speaking of automation means speaking of randomness, as we have seen, it is also true that DigiDan's answers to our questions can result from nothing but its randomness – there is a sense in which the definition of not being free as "not being physically constrained" can be nothing but random: resulting from nothing but "likely next 'tokens'", i.e. "sequences of characters that often co-occur in written text", is far from resulting from autonomous humans whose "will […] is a law to itself".

Thus, we may ask ourselves why we should trade Dennett and the experts in his work for DigiDan – and, more extensively, why

[5] See LSJ at https://www.perseus.tufts.edu/hopper/morph?l=automaton&la=greek#lexicon (accessed: May 2, 2024).

[6] See LSJ at https://www.perseus.tufts.edu/hopper/morph?l=automatizw&la=greek#lexicon (accessed: May 2, 2024).

[7] See LSJ at https://www.perseus.tufts.edu/hopper/morph?l=automatismos&la=greek#lexicon (accessed: May 2, 2024).

[8] See LSJ at http://stephanus.tlg.uci.edu/lsj/#eid=18224 (accessed: May 2, 2024).

[9] Especially from 1947 onwards, when the Ford Motor Company set up an automation department.

we should trade autonomy for automation[10]. Kant can continue to illuminate us. Whenever we trade autonomy for automation, which makes us as free as "the freedom of a turnspit", we can unburden ourselves from the burden of autonomy itself. Indeed, being autonomous means, first, that "moral law is possible" (Kant 1788, 5, 97), second, and consequently, that "imputation" (Kant 1788, 5, 97) is possible and, third, and consequently, that we are "culpable and deserving of punishment" (Kant 1788, 5, 100) whenever we (autonomously) fail. Conversely, being automated means, first, that "moral law" is replaced by a kind of "mechanism of nature" (Kant 1788, 5, 97) as nothing but a kind of "necessity of events in time" (Kant 1788, 5, 97), second, and consequently, that "imputation" is not possible and, third, and consequently, that we are not "culpable and deserving of punishment" whenever we fail – and, especially in our increasingly individualistic, competitive, accelerated and performing neoliberal societies, in which failure is almost unspeakable, unburdening ourselves from the increasingly unbearable burden of autonomy, which makes us "culpable and deserving of punishment" whenever we fail, may be precisely what we desperately try to do, sometimes consciously and sometimes unconsciously.

If we go back to the case of AI philosophers replacing human philosophers, we may think of DigiDan as an isolated case of testing. Yet, even though it is true that philosophy is markedly less characterised by the use of generative artificial intelligence than other areas typical of human creativity, it is also true that it may be a matter of time. On the one hand, the use of generative artificial intelligence exponentially affects professions from the least creative to the most creative, among which the various ways of writing are no exception, as several websites popularise[11]. On the other hand, the use of generative artificial intelligence exponentially affects students as future professionals who may naturalise it. As I myself can sense from my students as future creatives, if they are increasingly put under pressure by their increasingly individualistic, competitive, accelerated and performing educational environments, they may happen to increasingly use "design software [that] automates common processes. This allows you to save time. You can get designs […] quicker and enjoy a more efficient workflow. […] Its key feature is automation. […] [T]he software offers several tools for automating complicated processes. For example, it has

[10] I worked on the reasons why, from a philosophical perspective, we trade autonomy for automation in Chiodo 2022, 2023b and 2023c.

[11] For instance, see https://it.semrush.com/goodcontent/content-marketing-blog/ai-writing-tools/ (accessed: May 2, 2024).

an inbuilt library of stair and rail designs. […] You only need to enter a number into your variable to create the desired number of elements"[12]. Even in the case of the most creative (and, thus, pleasant) professions, if the pressure of overscheduling, overdoing and overperforming is taken to the extreme, technological automation, from totally standard "designs" to one's total alienation from one's (pleasant) creativity, may win out over the exercise of one's autonomy, i.e. autonomous creativity.

And, if it is true that even the most creative (and, thus, pleasant) professions are increasingly affected by the trade of human autonomy for technological automation, why should not it also happen to philosophers, who are increasingly asked to address the unprecedented complexity and uncertainty of the present time, from global health emergencies to global climate emergencies to global geopolitical crises to global economic crises to global crises of democracies and ideals to wars? More precisely, why should not it also happen to philosophers as humans who are exceedingly exposed to the increasingly unbearable burden of autonomously making sense of what increasingly seems to make no sense at all?

Yet, what AI philosophers lack, i.e. aesthetically living human existences, epistemically reflecting upon them and, finally, making sense of them, may also be what can save human philosophers, starting with aesthetically living human existences. Indeed, speaking of aesthetically living human existences means speaking not only of what underpins meta-sensemaking, as we have seen, but also of one of the most pleasant conditions of being human. Let us go back to our thought experiment according to which we struggle with the meaning of free will. Specifically, we are even faced with an existential crisis, tormented, in need of both making the right decision and living up to the exercise of free will and confused – which means precisely that we live aesthetically: any of our epistemic questions correlates with what we may define as our aesthetic questions, in that speaking of struggling, existential crisis, being tormented, being in need and being confused means speaking of sensations, emotions and feelings. And speaking of sensations, emotions and feelings means speaking of what can push us to give sensible answers even more than our epistemic reasons for answering our questions – our aesthetic reasons for answering our questions can be even stronger than our epistemic reasons. Whenever Dennett and the experts in his work, as humans, answer questions on free will, they can have both epistemic

[12] See https://academy.archistar.ai/top-ten-design-software-for-architects (accessed: May 2, 2024).

and aesthetic reasons for giving the most sensible answers they can, in that they know both what it means to be epistemically challenged and, especially, what it means to be aesthetically challenged by even torturing sensations, emotions and feelings. Once Dennett and the experts in his work, as humans, answer questions on free will, they can also make sense of their both epistemic and aesthetic lives and, thus, find a sense of resolution and even evolution, which works as an essential push to give the most sensible answers possible. Conversely, whenever DigiDan, as technology, answers questions on free will, it cannot have epistemic and, especially, aesthetic reasons for giving the most sensible answers it can, in that it does not know what it means to be epistemically challenged and, especially, what it means to be aesthetically challenged by even torturing sensations, emotions and feelings. Once DigiDan, as technology, answers questions on free will, it can do nothing but automatically "predict likely next 'tokens'" – and automatically "predict likely next 'tokens'" has nothing to do with making sense of both epistemic and aesthetic lives and, thus, finding a sense of resolution and even evolution: in the case of DigiDan, as technology, speaking of a sense of resolution and even evolution as an essential push to give the most sensible answers possible means speaking of nothing but a nonsense.

If my argument makes sense, we may try to consider at least the following two last remarks: the first is on a strictly philosophical issue and the second is on an issue between philosophy and technology policy.

First, we should stop thinking of technologies as kinds of anthropomorphic artefacts, which means not only that they cannot always be designed according to human prerogatives (see at least Sætra 2022), starting with human somatics, but also that they cannot always be used according to human perspectives, starting with human epistemic and, especially, aesthetic challenges. Conversely, we should intensify our philosophical work on the reasons why we are exponentially willing to trade the prerogatives of being human, i.e. autonomous, for the prerogatives of being technological, i.e. automated. If my argument makes sense, the starting hypothesis may be that what we are exponentially willing to trade is the unbearable burden of human autonomy for what can unburden us from it, i.e. technological automation – what we are exponentially willing to trade is human autonomy, as what is unbearably challenged by the unprecedented complexity and uncertainty of the present time, for technological automation, as what can unburden us from making sense of them, starting with our even torturing

sensations, emotions and feelings, i.e. our aesthetic lives, whenever we desperately fall into struggle, existential crisis, torment, need and confusion.

Second, work on technology policy should also be intensified. If it is true that AI philosophers cannot replace human philosophers, in that they lack essential conditions for giving us sensible answers, it is also true that we, as citizens, should always be put in a position to know what is the source of the answers to our questions. We cannot trust DigiDan (and the caged parrot) as a reliable source of knowledge, as we have seen. But, even though we decide to trust it for a variety of reasons, starting with the kind of desperate trade argued above, we should always have the right to know what is the source of the answers to our questions, in that it makes a big difference (in the future, if not in the present), as we have seen. Europe is well on its way to guarantee it[13], but more work is needed when it comes to education, starting with students exponentially affected by the use of generative artificial intelligence, who may become future professionals who may naturalise it, as we have seen. More precisely, when it comes to education, we, as (human) philosophers, may try to teach our students that meta-sensemaking may be far better than randomness even when it means desperately falling into struggle, existential crisis, torment, need and confusion, in that the former can give us what the latter cannot give us: making us not only epistemically live but also, and especially, aesthetically live – making us capable of sensations, emotions and feelings, and not atrophied, and, thus, making us capable of both the pleasure of feeling pleasure and the pleasure of feeling in itself even when we do not feel pleasure, but struggle, existential crisis, torment, need and confusion (from which, differently from DigiDan, we can learn essential lessons on what it means to be autonomous humans).

References

Anscomb, C., 2021, "Creative agency as executive agency. Grounding the artistic significance of automatic images". *The journal of aesthetics and art criticism* 79/4: 415-427.

Arielli, E., 2021, "Extended aesthetics. Art and artificial intelligence", *Proceedings of the European society for aesthetics* 13: 1-3.

Barale, Alice, ed., 2020, *Arte e intelligenza artificiale. Be My GAN*, Milano: Jaca Book.

[13] See https://artificialintelligenceact.eu/ (accessed: May 6, 2024).

Boden, M.A., 2010, "The Turing test and artistic creativity", *Kybernetes* 39/3: 409-413.
Boden, M.A., 2019, *From fingers to digits. An artificial aesthetic*, Cambridge: MIT Press.
Cetinic, E., She, J., 2022, "Understanding and creating art with AI. Review and outlook", *ACM transactions on multimedia computing, communications and applications* 18/2: 1-22.
Chiodo, S., 2022, "Human autonomy, technological automation (and reverse)", *AI & society* 37/1: 39-48.
Chiodo, S., 2023a, "Intelligenza artificiale e destino della forma", *Scienza e filosofia* 30/2: 210-234.
Chiodo, S., 2023b, *Technology and the overturning of human autonomy*, Cham: Springer.
Chiodo, S., 2023c, "Trading human autonomy for technological automation", in S. Lindgren, ed., *Handbook of critical studies of artificial intelligence*, Cheltenham-Northampton: Edward Elgar Publishers: 67-78.
Chiodo, S., 2024, *Predicted humans. Emerging technologies and the burden of sensemaking*, Abingdon-New York: Routledge-Taylor & Francis Group.
Coeckelbergh, M., 2017, "Can machines create art? *Philosophy & technology* 30/3: 285-303.
Elgammal, A., 2019, "AI is blurring the definition of artist. Advanced algorithms are using machine learning to create art autonomously", *American scientist* 107/1: 18-22.
Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M., 2017, "CAN. Creative adversarial networks, generating 'art' by learning about styles and deviating from style norms", *Arxiv*: 1706.07068.
Fazi, M.B., 2019, "Can a machine think (anything new)? Automation beyond simulation", *AI & society* 34: 813-824.
Goodfellow, I.J., 2016, "NIPS 2016 tutorial. Generative adversarial networks", *Arxiv*: 1701.00160.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014, "Generative adversarial nets", *Advances in neural information processing systems* 27: 2672-2680.
Kant, I., 1785, in Gregor, M.J., ed., *Groundwork of the metaphysics of morals*, Cambridge: Cambridge University Press.
Kant, I., 1788, in Gregor, M.J., ed., *Critique of practical reason*, Cambridge, Cambridge University Press.
Liddell, H.J., Scott, R., 1940, *A Greek-English lexicon* (LSJ), Oxford: Clarendon Press.

Manovich, L., 2019, Defining AI art: three proposals. AI and dialogue of cultures, Saint-Petersburg: Exhibition catalogue.
Mazzone, M., Elgammal, A., 2019, "Art, creativity, and the potential of artificial intelligence", *Arts* 8/1: 26.
Miller, A.I., 2019, *The artist in the machine. The world of AI-powered creativity*, Cambridge: MIT Press.
Rifkin, J., 1995, *The end of work. The decline of the global labour force and the dawn of the post-market era*, New York: Putnam.
Sætra, H.S., 2022, "Robotomorphy", *AI & ethics* 2: 5-13.
Schwitzgebel, E., Schwitzgebel, D., Strasser, A., 2023, "Creating a large language model of a philosopher", *Arxiv*: 2302.01339.
Steiner, S., 2017, "Art. Brought to you by creative machines". *Philosophy & technology* 30/3: 267-284.
Strasser, A., Crosby, M., Schwitzgebel, E., 2023, "How far can we get in creating a digital replica of a philosopher?", in R. Hakli, P. Makela, J. Seibt, eds., *Social robots in social institutions. Proceedings of "Robophilosophy 2022*", Amsterdam: OIS Press, 371-380.
Taylor, G.D., 2014, *When the machine made art. The troubled history of computer art*, New York: Bloomsbury.
Terzidis, K., Fabrocini, F., Lee, H., 2022, "Unintentional intentionality. Art and design in the age of artificial intelligence", *AI & society* 38: 1715-1724.
Turing, A.M., 1950, "Computing machinery and intelligence", *Mind* 59: 433-460.