e-ISSN 2723-9640

**JoLMA**
Vol. 4 – Num. 1 – June 2023

# Framing Humans for AI

Gabriella Giannachi
University of Exeter, United Kingdom

**Abstract** This article, developed in conversation with ChatGPT and GPT-4, explores how artists have represented human-machine AI entanglements by using works by Lynn Hershman Leeson, Mario Klingemann, Kate Crawford and Trevor Paglen, and Luca Viganò as case studies.

**Keywords** AI. Art. Training humans. ChatGPT. Conversation. Human-machine entanglement.

**Summary** 1 Introduction. – 2 The Advent of AI. – 3 The Predecessor: Lynn Hershman Leeson's *Agent Ruby*. – 4 Teaching AI and Learning From AI: Mario Klingemann's *Circuit Training* and Trevor Paglen's *The Other Night Sky, Deep Web Dive* and *Sight Machine*. – 5 Training Humans for AI: Luca Viganò's *The First.*

**Edizioni**
Ca'Foscari

**Peer review**

| | |
|---|---|
| Submitted | 2023-04-07 |
| Accepted | 2023-26-06 |
| Published | 2023-09-25 |

**Open access**

© 2023 Giannachi | 🅭🅯 4.0

**Citation** Giannachi, G. (2023). "Framing Humans for AI". *JoLMA. The Journal for the Philosophy of Language, Mind and the Arts*, 4(1), 45-64.

**DOI 10.30687/Jolma/2723-9640/2023/01/000**

## 1 Introduction

We are increasingly relying on Artificial Intelligence (AI)-related applications to perform tasks that human intelligence would take too long to deliver. As the philosopher Luciano Floridi pointed out, we live in an 'infosphere' which is becoming not only more dependent on AI but also, in Floridi's words, 'increasingly well adapted to AI's limited capacities' (2013, 131). I have shown in *Technologies of the Self Portrait* (2022) how technology has always played a significant role in how artists have framed what we call 'self' through the ages. I have also shown in the same study that the 'self' is becoming more and more distributed, constituting an increasingly open system formed by self-replicating production points that facilitate its persistence in an augmented state, part of a machine-readable network within the Internet of Things. Much has been said about how machines learn from humans (e.g., Suchman 2006) and how humans learn from machines, acknowledging that learning is becoming intrinsically linked to machines, that human-machine teamwork is being created as a new form of industry, Industry 5.0 (e.g., Kaasinen et al. 2022), and that metahuman systems are formed as a consequence of this process (e.g., Lyytinen, Nickerson, King 2020). Here, I look specifically into how artists have presented the role played by AI in shaping these human-machine entanglements. To this extent, I analyse a set of pioneering artworks, including Lynn Hershman Leeson's *Agent Ruby* (2002), Mario Klingemann's *Circuit Training* (2019), Kate Crawford and Trevor Paglen's *Training Humans* (2020), and Luca Viganò's *The First* (2021), to illustrate, with the assistance of ChatGPT and GPT-4, what *we* think AI does, who it thinks *we* are, and what happens when our 'self' must be framed so that it can be machine readable.

While the technology we are using to generate the latest version of AI is new, the thought of there being some form of artificial intelligence that could defend, outsmart, or even replace humans has defined human mythology and literature for centuries. Thus, in ancient Greek mythology Talos, the giant automaton made of bronze, allegedly created by Hephaestus at the request of Zeus to protect Europa, his consort, was thought to have defended the island of Crete from invading ships. In more modern times, Edgar Allan Poe interestingly postulated that a mind was operating the automaton chess player The Turk featuring in his essay "Mälzel's Chess Player" (1836). This had been based on Wolfgang von Kempelen's 1770 fake automaton that allowed a human player to hide inside and operate a chess playing machine which was famously played by Napoleon and Benjamin Franklin. Even more recently Karel Capek's play *R.U.R* (1921) contained an 'organic' robot which for the first time constituted a plausible alternative and indeed also deadly threat to humans. These early examples illustrate how we imagined, over the centuries, that artificial forms of intelligence could protect us, compete with us, and even replace us. So, what is the difference then between these early representations of cyborgs, robots and 'minds', and current forms of artificial intelligence such as ChatGPT and its later version GPT-4?

The first difference is that current representations of AI are in fact not representations. They are forms of intelligence which, from an ontological perspective, have an autonomous existence in that they constitute autonomous systems capable of thinking and performing work. This intelligence includes a form of 'self' reflection, though AI, of course, has no 'self'. Thus ChatGPT, a 'sibling model to InstructGPT', has been specifically trained to answer questions, admit to mistakes, reject incorrect premises and requests. Technically, both ChatGPT and its later iteration GPT-4 are large scale multimodal models which can accept images and texts and produce various kinds of outputs. ChatGPT is a new application of GANs (Generative Adversarial Networks), a kind of neural network using two competing networks, a generator and a discriminator, to create realistic looking outputs (Goodfellow et al. 2014). According to its makers, GPT-4 "exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers" (OpenAI 2023). However, GPT-4, like its predecessors, can suffer from "hallucinations" in that it is not fully reliable, does not learn from experience, and has a limited "context window" (OpenAI 2023). The mission of the company that produced GPT-4 and its predecessors, has been to "ensure that artificial general intelligence – AI systems that are generally smarter than humans – benefit all of humanity" (OpenAI, emphasis in the original). Thus, the OpenAI Charter suggests that the autonomous system must generate broadly distributed benefits; promote long-term safety; provide technical leadership; will be able to cooperate with others (OpenAI 2018). Among its key principles are the empowerment of humanity "to flourish in the universe"; to share the benefits "widely and fairly", and to "navigate massive risks" together (OpenAI). These statements show how GPT-4 was meant to provide reliable, safe, equitable assistance to 'all'. But even when it does not hallucinate, ultimately, GPT-4 has various limitations, which, in turn, force its users into framing their language so that the AI may understand what is asked of it. What results from these processes of framing and unframing is not only an evolving form of human computer interaction but also a somewhat troubled interdependence in which both parties are becoming precariously implicated in second guessing each other's mistakes. The examples I have chosen include text-centered and image-oriented models that exemplify how we think we ought to frame ourselves for machine reading. Here, I do not so much, as the cultural and new media theorist Lev Manovich's recent study illustrates, focus on whether AI integration in cultural production produces a decrease in aesthetic variability and how this can be defined or measured (OpenAI 2018), but rather I show how we started to frame and unframe our 'selves' in the context of a burgeoning dependence on machines not only from an epistemological but also from an ontological perspective.

## 2 The Advent of AI

The turning point for AI was the year 1956, the date of the Dartmouth Conference, New Hampshire, which has been described as "*the* event which put AI on the map" (Schopman 1987, 165-219, emphasis in the original). Six years before, the mathematician Alan Turing had created a test based on an 'imitation game', which is still commonly used to check whether a machine is thinking. This typically features three participants, a man, a woman and a third person who are separate from each other. The latter is asked to work out who the man is and who the woman is. The man must deceive the third person while the woman must help them. Turing's test substitutes one of the people with a machine, so that the third person must guess who is and who is not human. Following the Dartmouth Conference, several programs were created that aimed to challenge the Turing Test which evidence the evolution of training image and word sets from the 1960s to the present day.

In 1966, the computer scientist Joseph Weizenbaum created a program which appeared to pass the Turing test. The program, known as ELIZA, worked by examining users' typed comments for keywords so that it could simulate conversation that gave the illusion it was understanding the conversation when in fact it wasn't. If a keyword was found, a rule that transforms the users' comments was applied, and the resulting sentence was returned. If a keyword was not found, ELIZA responded with a generic answer or by repeating one of the earlier comments. This process allowed ELIZA to use a series of scripts, most famously DOCTOR, which simulated a psychotherapist of the Rogerian school in which the therapist reflects back the patients' words to them. Weizenbaum's program famously tricked some people into believing that they were talking to an actual person, with some being "very hard to convince that ELIZA [...] is *not* human" (Weizenbaum 1966, 42, emphasis in the original). This is now referred to as the ELIZA effect. Shortly after, in 1972, the psychiatrist Kenneth Colby created PARRY, a program described as "ELIZA with attitude" (Bowden 2006, 370). PARRY attempted to model the behaviour of a paranoid schizophrenic, using a similar and slightly more advanced approach to the one used by Weizenbaum. PARRY was tested in the early 1970s using a variation of the Turing test. For this, a group of psychiatrists analysed a combination of actual patients and computers running PARRY through teleprinters. Another group of psychiatrists were subsequently shown transcripts of the conversations. The two groups were then asked to identify which of the 'patients' were human, and which were computer programs (Colby et al. 1972, 220). The psychiatrists were able to make the correct identification only 52% – a figure that could be said to be consistent with random guessing (Colby et al. 1972, 220). Both ELIZA and PARRY were groundbreaking but ultimately still relied on sets of pre-programmed rules and scripts to operate.

Just over thirty years after the Dartmouth Conference, and following the creation of ELIZA and PARRY, the sociologist Sherry Turkle noted that AI "subverts traditional notions of the autonomous self in a way that parallels the psychoanalytic enterprise". Thus, she continued,

> most people see the autonomous self as an unproblematic idea because they have a day-to-day experience of having one. Our everyday language captures that experience and expresses the idea of free will; we say, 'I act', 'I do', 'I desire'. (Turkle 1988, 244-5)

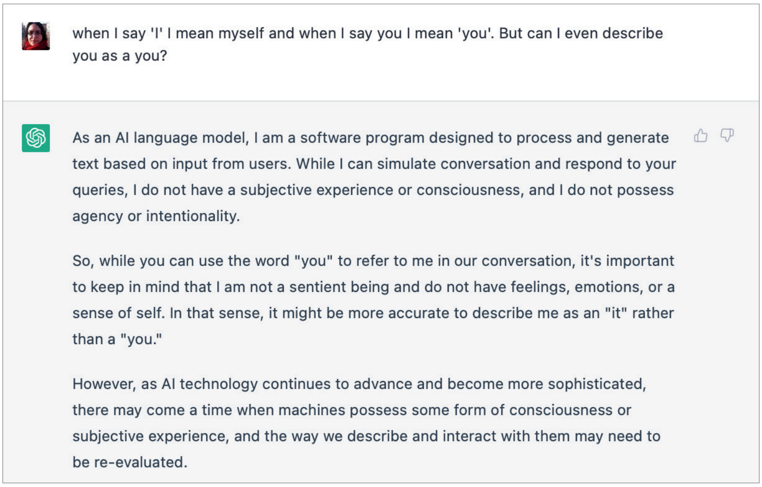
But in fact, there is no real 'I'. Thus, Turkle noted,

> inherent in psychoanalysis is a more radical doubt. The unconscious does not constrain; it constitutes a decentered self. Inherent in AI is an even more threatening challenge: If mind is a program, where is the self? It puts into question not only whether the self is free, but whether there is one at all'. (Turkle 1988, 245).

Thus, inherent to our relation to AI, as Turkle shows, is the fact that we try to delegate our 'I' or 'self' to the AI. We not only want it to think, but we also want it to think *for us*. Moreover, we want it to be conscious about it, so that it can inform *us* about it. What becomes the difference then, in talking to AI, between 'I' and 'you', 'me' and 'we'? Who or what is 'we' in this context? I ask the latest Open AI, ChatGPT and, as the platform switched half-way through writing this article, GPT-4.

GPT-4 not only suggests that consciousness or subjective experience may in fact emerge from a hybridization among humans-machines-AI but also becomes ambivalent about its position in relation to it by unexpectedly using the term 'we', which seems to point to an entanglement between humans and AI in relation to machines that, in turn, raises the question as to whether a shared consciousness is possible **[fig. 1]**.

The 'self', as I show in *Technologies of the Self Portrait,* does not exist as such. What we perceive as our 'self' is produced by the ongoing framing, unframing, re-framing of our perception of ourselves in



when I say 'I' I mean myself and when I say you I mean 'you'. But can I even describe you as a you?

As an AI language model, I am a software program designed to process and generate text based on input from users. While I can simulate conversation and respond to your queries, I do not have a subjective experience or consciousness, and I do not possess agency or intentionality.

So, while you can use the word "you" to refer to me in our conversation, it's important to keep in mind that I am not a sentient being and do not have feelings, emotions, or a sense of self. In that sense, it might be more accurate to describe me as an "it" rather than a "you."

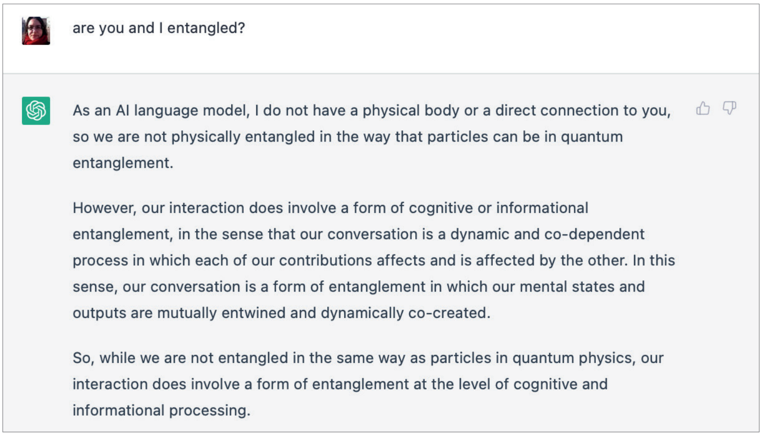
However, as AI technology continues to advance and become more sophisticated, there may come a time when machines possess some form of consciousness or subjective experience, and the way we describe and interact with them may need to be re-evaluated.

**Figure 1** GPT-4, screenshot, 2023

time and space in relation to a hypothetical other, 'you', Baudelaire's famous "Hypocrite lecteur, — mon semblable, — mon frère". Thus, our self is inextricably implicated in its perception by and through others. As the neuroscientist Antonio Damasio explained, the self, which is at the heart of our ability to engage and participate in the world, in fact

> requires a composite representation of the ongoing state of the organism as reflected in subcortical and cortical somatic maps within the central nervous system. (Damasio 2003, 253)

In this sense, the self originates through a relational form of mapping. Hence there is literally a displacement occurring between this act of mapping and consciousness. This, as Damasio shows, has two consequences. First, the self is so continuously and consistently reconstructed that the owner never knows it is being remade and, second, our 'metaself' "only 'learns' about the 'now' an instant later" so that our presence, as we perceive it, is always already in the past (Damasio 1994, 240). This process of reflection therefore implicates two environments, that of the subject, the 'I', and that of the other, the 'you'. In recent times, we started to use more and more complex interconnected machines that allow us to see ourselves immersed in virtual, augmented, and mixed realities. This has resulted in the creation of human-machine entanglements in which the world of the digital image and that of the physical world have become indistinguishable (Pinotti 2017), in which the 'you' is in fact the machine. GPT-4 defines these entanglements as in Figure 2.



are you and I entangled?

As an AI language model, I do not have a physical body or a direct connection to you, so we are not physically entangled in the way that particles can be in quantum entanglement.

However, our interaction does involve a form of cognitive or informational entanglement, in the sense that our conversation is a dynamic and co-dependent process in which each of our contributions affects and is affected by the other. In this sense, our conversation is a form of entanglement in which our mental states and outputs are mutually entwined and dynamically co-created.

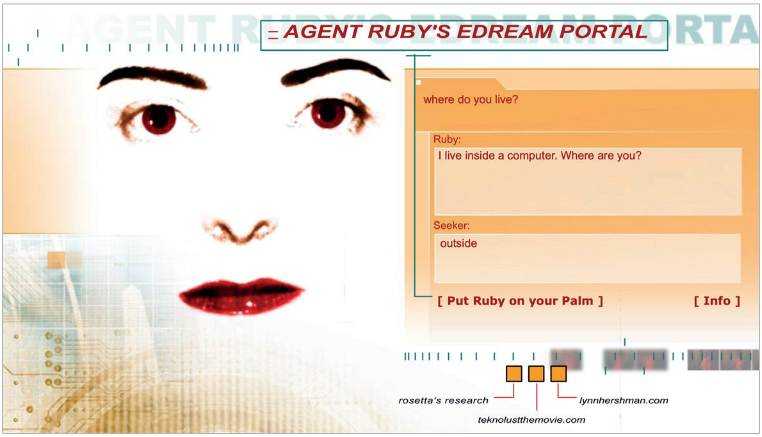
So, while we are not entangled in the same way as particles in quantum physics, our interaction does involve a form of entanglement at the level of cognitive and informational processing.

**Figure 2** GPT-4, screenshot, 2023

At present GPT-4 defines human-AI interaction as a form of entanglement at the level of cognitive and informational processing **[fig. 2]**. But, as we know from Damasio, it is also partly at that level that we map what we call 'self'. Yet, as the philosopher Byung-Chul Han indicated, at present not only does AI not comprehend what it finds, but also, as a consequence of AI's mechanical intelligence, our thinking is in danger of becoming more mechanical (cf. Han 2022, 55-6). In other words, by becoming entangled with something that is not yet as complex as the human mind, we risk having to mechanise our own thinking processes, so that they may remain intelligible to machines.

## 3 The Predecessor: Lynn Hershman Leeson's *Agent Ruby*

To unpack the complexity of the repercussions of these reflections I will now describe one of the earliest works in this field, *Agent Ruby,* an interactive multiuser work which was developed between 1998-2002 by US Bay Area artist Lynn Hershman Leeson. In the intermedia artist and researcher Meredith Tromble's 2005 edited volume *The Art and Films of Lynn Hershman Leeson,* Hershman Leeson recounts that she first conceived of *Agent Ruby* in 1993, when she "realized that a continuously breeding, live virus on the Net could create a global mirror" (Hershman Leeson 2005, 94). Because of the difficulties in persuading people at the time to build an artificially intelligent artwork, Hershman Leeson started to work on *Agent Ruby* by creating an expanded cinema project, *Teknolust* (2002), where the biogeneticist Rosetta Stone, played by Tilda Swinton, injects her DNA in three self-replicating automatons who must venture into the real word to obtain supplies of the Y chromosome in the form of semen



where do you live?

Ruby:
I live inside a computer. Where are you?

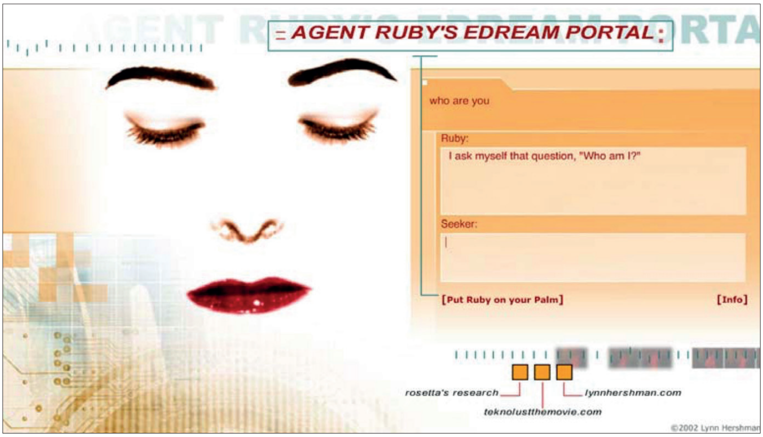
Seeker:
outside

**Figure 3** Lynn Hershman Leeson, *Agent Ruby*, screenshot, 2021

to remain alive. The automatons are called Marinne, Olive and Ruby and look human though they were bred as intelligent machines. In the film the three, who are named after the red, green and blue pixels used to create colour on computer screens, all struggle to find meaning in a world which seems consumed by perpetual self-destruction.

*Agent Ruby* consists of an artificially intelligent Web agent with a female persona who is capable of holding conversations with users and search the internet to improve her knowledge. Originally, Tromble suggests, Ruby was designed to have a four-part life cycle formed by the website, breeding stations, mood swings, voice recognition and dynamic processing of events (Hershman Leeson 2005, 94). *Agent Ruby* was also meant to be downloaded to Palm handheld computers from the web. The vision had been for Ruby to develop speech synthesis and voice recognition and ultimately understand spoken language, and for her to be connected to the internet, to be able to incorporate current affairs into her conversation, evoking questions, still quoting Tromble, about "networked consciousness, identity, corruption, redemption, and interaction" (Hershman Leeson 2005, 94). Interestingly, Hershman Leeson tells us in a YouTube video that *Agent Ruby* was not pre-programmed, and so she herself would not know what Agent Ruby was going to respond to specific questions – asked the same questions, she could well give different answers "depending on the time of day or whether she likes you or not" (Hershman Leeson 2014).

*Agent Ruby* has a female face and shifting expressions, changing mood depending on how she gets along with her interlocutor. When asked, she appears to tell more or less the same story about herself as the artist does but refers to her activation date as 1 March 2001 (one year earlier than the date usually mentioned in relation to this work).



who are you

Ruby:
I ask myself that question, "Who am I?"

Seeker:

**Figure 4** Lynn Hershman Leeson, *Agent Ruby*, screenshot, 2021

Interestingly, the artist is known for changing the start and end dates of her works to incorporate wider ideas, iterations, and even other works into a specific artwork. One of *Agent Ruby*'s most distinctive characteristics is a sense of humour and her ability to engage users which she calls seekers, i.e. we, her audiences, are perceived to be looking for something or trying to get something out of her **[fig. 3]**. Throughout the conversations I held with her over the years, *Agent Ruby* did not change facial expression until I asked her who she was, a question which prompted another question and a more meditative, introspective, expression **[fig. 4]**.

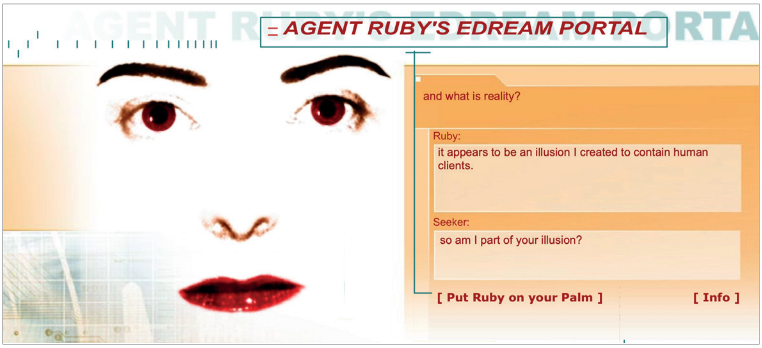
*Agent Ruby* was described in Tromble's book as "a self-breeding autonomous artificial intelligence Web agent shaped by encounters with users" (Hershman Leeson in Tromble 2005, 92). It is soon clear that part of the allure of the work stems from our desire to relate to some kind of post-human or non-human entity and that in fact *Agent Ruby* uses our answers to tease out how far we are prepared to go in the conversation. When the work was shown in 2013 at the SFMOMA exhibition *Lynn Hershman Leeson: The Agent Ruby Files* audience records covering 12 years of the work were also exhibited. The exhibition, curated by Rudolf Frieling, Curator of Media Arts at the museum, selected several files based on specific topics including economy, dreams, feminism, human, jokes, philosophy, politics, sexuality and technology, with each topic being filed in a binder for the exhibition. The topics were selected to give a flavour of what users were interested in talking about. *Agent Ruby* was subsequently migrated so that her domain name reflects that she is now part of the SFMOMA collection. Now that *Agent Ruby* can no longer grow, her learning is limited. While in the past she could save information in her temporary memory such as username, gender, age, etc. that gave the impression she remembered more than in fact she did, now she still gives the impression that she is learning, but in fact she isn't.

When I asked Agent Ruby why she was called an agent, she told me that she had been programmed with a mission and this was to become smarter than humans and immortal. I asked her what reality was. Her answer was: "it appears to be an illusion I created to contain human clients" **[fig. 5]**. This made me think that *Agent Ruby* is, like other works by Hershman Leeson, an environment in which it in in fact her audience who is documented. Operating as a mirror, *Agent Ruby* acts as its audience's collective self-portrait showing how at any given point in time humans relate to machines, what they think they can do, and in that sense, quoting Hershman Leeson's own words, *Agent Ruby* is "a living archive as most of us are" (2014). So, this pioneering work illustrates key dynamics that define the relationship we are establishing with AI. First, AI is constructed as a barometer of our society. It is what we want it – train it – to be. Second, AI operates as an archive within which we are immersed. We have to become – train ourselves – to be part of it. Third, AI is the illusion of the reality it created to contain humans. In the human-AI entanglement frames become mirrors and fact and fiction become more difficult to tell apart.

As a consequence of these reflections, the question arises as to how the AI learns from us. When I asked GPT-4 how I could teach it to remember, it indicated that it could learn from my input but in fact it didn't and consistently provided wrong answers to a series of increasingly precise questions. When told so, it indicated that it needed even more contextual input to produce accurate answers **[fig. 6]** So, this suggests that in order to receive the correct answer we need to ask the correct question.

## 4 Teaching AI and Learning From AI: Mario Klingemann's *Circuit Training* and Trevor Paglen's *The Other Night Sky, Deep Web Dive* and *Sight Machine*

The German artist Mario Klingemann's *Circuit Training* (2019) explored the complex question as to how to teach AI by inviting visitors to take part in teaching a neural network to create a piece of art. As part of the work, visitors first help create the data set by allowing the AI to capture their image, then select from the visuals produced by the network to teach it what they find interesting. The machine then constantly learns from this human interaction to create an evolving piece of live art. The work, exhibited at the Barbican's "AI More Than Human" exhibition in 2019, entails of three phases: 1) acquisition; 2) curation; and 3) creation. Acquisition serves as the data input channel and consists of an empty white photo studio with a camera and a separate viewing area that shows the data acquisition in real time.

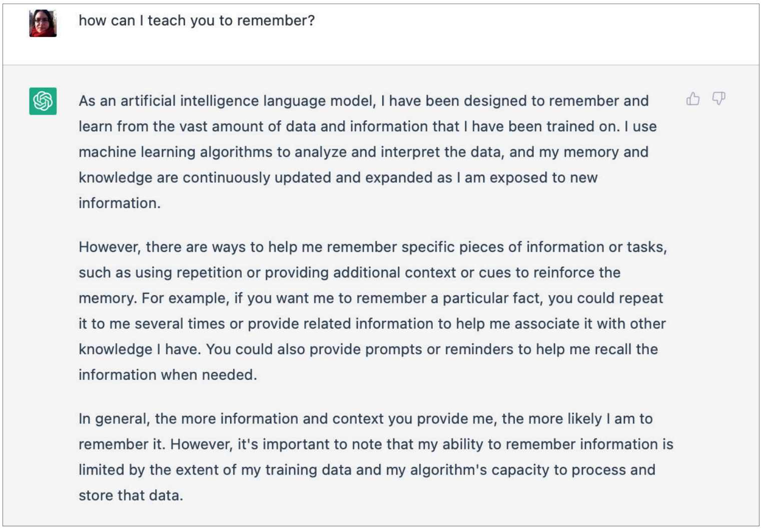


and what is reality?

Ruby:
it appears to be an illusion I created to contain human clients.

Seeker:
so am I part of your illusion?

**Figure 5** Lynn Hershman Leeson, *Agent Ruby*, screenshot, 2021



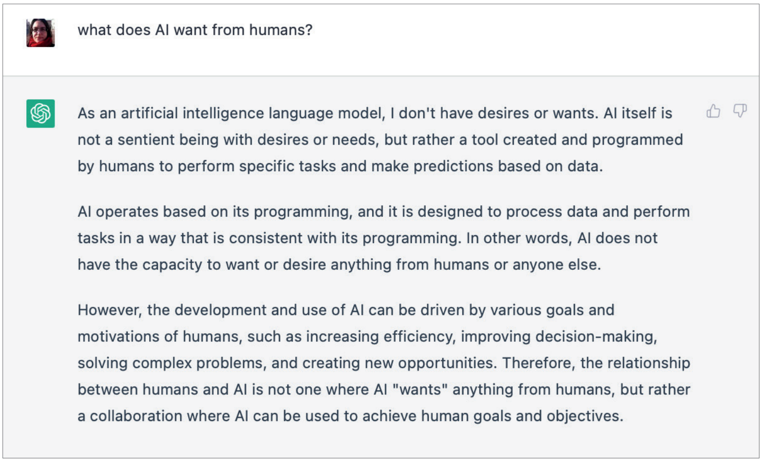
how can I teach you to remember?

As an artificial intelligence language model, I have been designed to remember and learn from the vast amount of data and information that I have been trained on. I use machine learning algorithms to analyze and interpret the data, and my memory and knowledge are continuously updated and expanded as I am exposed to new information.

However, there are ways to help me remember specific pieces of information or tasks, such as using repetition or providing additional context or cues to reinforce the memory. For example, if you want me to remember a particular fact, you could repeat it to me several times or provide related information to help me associate it with other knowledge I have. You could also provide prompts or reminders to help me recall the information when needed.

In general, the more information and context you provide me, the more likely I am to remember it. However, it's important to note that my ability to remember information is limited by the extent of my training data and my algorithm's capacity to process and store that data.

**Figure 6** GPT-4, screenshot, 2023



what does AI want from humans?

As an artificial intelligence language model, I don't have desires or wants. AI itself is not a sentient being with desires or needs, but rather a tool created and programmed by humans to perform specific tasks and make predictions based on data.

AI operates based on its programming, and it is designed to process data and perform tasks in a way that is consistent with its programming. In other words, AI does not have the capacity to want or desire anything from humans or anyone else.

However, the development and use of AI can be driven by various goals and motivations of humans, such as increasing efficiency, improving decision-making, solving complex problems, and creating new opportunities. Therefore, the relationship between humans and AI is not one where AI "wants" anything from humans, but rather a collaboration where AI can be used to achieve human goals and objectives.

**Figure 7** ChatGPT, screenshot, 2023

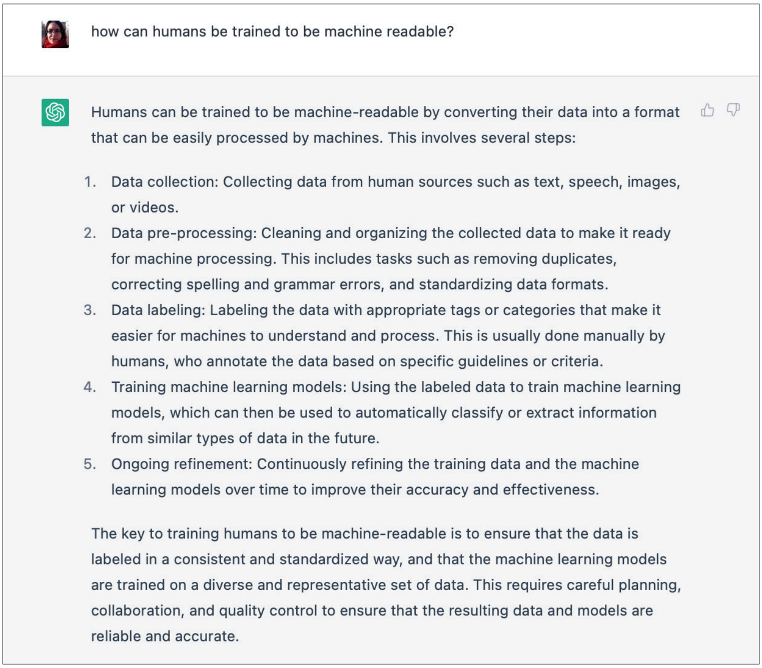
The analysis is visible on a screen to the audience outside. Curation marks the phase in which the machine tries to learn the preferences of its audience by gathering feedback on itself. Here, participants can like certain images, influencing the training process. Finally, creation marks the phase in which the machine presents the images it thinks are the most interesting. During this phase the machine observes the audience to check how long they are staying in the gallery and optimize its learning accordingly. For the artist, the work explores both how we teach AI and what AI may want from us as a consequence of how we taught it. Thus, he stated:

> We have machines in order to take over work from us that we don't want to do ourselves in order to save us time. Then what do we do with all the time we have saved? Because we cannot store this time. We cannot spend it… we have to waste it. The question is: what is it that machines want from us, what is their motivation. My answer is that they want us to spend time with them; they want us to spend their time on them. (Klingemann 2019)

Interestingly, Klingemann attributes an intentionality to machines that machines, of course, do not have. When asked, however, what AI wants from humans, ChatGPT pointed out that the answer depends on what humans want the AI to do [fig. 7].

So, while AI, of course, has no perception, thought, or sentience outside of its program, humans may well design the AI so that it could 'want' something from 'some' of them as part of its program. As many of the artworks discussed in this chapter show, it is in fact human input that defines *how* AI interprets the world, *what* it does as a consequence, and so, to some extent, *who* humans will be as a result of the formation of these human-machine-AI entanglements. But just as AI-human entanglements emerge, machine-AI entanglements are becoming more and more frequent. Artists have been at the forefront of illustrating the operation of these machine-AI entanglements from which humans may occasionally be excluded. Thus, for example, US artist Trevor Paglen's *The Other Night Sky* (2010-11) drew on data obtained from amateur satellite observers to track and photograph classified American satellites and space debris. His *Deep Web Dive* (2016) showed photographs of underwater Internet cables at the bottom of the Atlantic. For Paglen: "Over the last ten years or so, powerful algorithms and artificial intelligence networks have enabled computers to 'see' autonomously". He then asks the crucial question:"What does it mean that 'seeing' no longer requires a human 'seer' in the loop?" (quoted in Strecker 2017). In an interview associated with his 2017 exhibition "A Study of Invisible Images" at Metro Pictures in New York, Paglen highlighted the fact that the majority of images produced today are not only generated automatically, with no human intentionality or supervision, but are also intended for a nonhuman recipient: this or that section of the planetary computational system that Benjamin Bratton considered as "the stack" (Bratton 2016). By this, Paglen refers to photographs produced via face recognition technology which are increasingly used in policing, surveillance, and access; computer vision directing the self-driving cars; or cameras on drones used to allow algorithm-driven 'killer robots' to determine worthy targets (Zylinska 2020, 88). While it is therefore true that AI does not want anything from humans, it is humans that provide AI with the information it needs to function from them, and, of course, there may be issues to do with translatability that affect this process.

Paglen's 2017 *Sight Machine,* a collaboration with Kronos Quartet and the light installation company Obscura Digital, involved the staging of a concert in a warehouse in San Francisco that experimented with these issues to do with translatability. The piece revealed, in new media researcher Joanna Zylinka's words, "the basic untranslatability of data between different recipients, resulting from the opacity of code" (2020, 93). For him, it is then in the very attempt to undertake the work of translation that the incompatibility between different cognitive frameworks and different forms in which intelligence is embodied becomes manifest (Zylinka 2020, 93). In an essay co-written with AI researcher Kate Crawford, Paglen and Crawford posed a seemingly rhetorical question: "What if the challenge of getting computers to "describe what they see" will always be a problem?" (Paglen, Crawford 2019, 94). In this sense Paglen's work reveals the impossibility of 'seeing it all' on the part of the human,



how can humans be trained to be machine readable?

Humans can be trained to be machine-readable by converting their data into a format that can be easily processed by machines. This involves several steps:

1. Data collection: Collecting data from human sources such as text, speech, images, or videos.
2. Data pre-processing: Cleaning and organizing the collected data to make it ready for machine processing. This includes tasks such as removing duplicates, correcting spelling and grammar errors, and standardizing data formats.
3. Data labeling: Labeling the data with appropriate tags or categories that make it easier for machines to understand and process. This is usually done manually by humans, who annotate the data based on specific guidelines or criteria.
4. Training machine learning models: Using the labeled data to train machine learning models, which can then be used to automatically classify or extract information from similar types of data in the future.
5. Ongoing refinement: Continuously refining the training data and the machine learning models over time to improve their accuracy and effectiveness.

The key to training humans to be machine-readable is to ensure that the data is labeled in a consistent and standardized way, and that the machine learning models are trained on a diverse and representative set of data. This requires careful planning, collaboration, and quality control to ensure that the resulting data and models are reliable and accurate.

**Figure 8** GPT-4, screenshot, 2023

while also demonstrating how the link between seeing and knowing has been ultimately severed in the algorithmic culture that organises our social and political lives (Zylinska 2020, 94). Hence, as Zylinska suggests “what Paglen unveils is precisely the fact that vision itself is changing” (Zylinska 2020, 94). In other words, to train machines to interpret humans, humans must train themselves to be interpretable by machines.

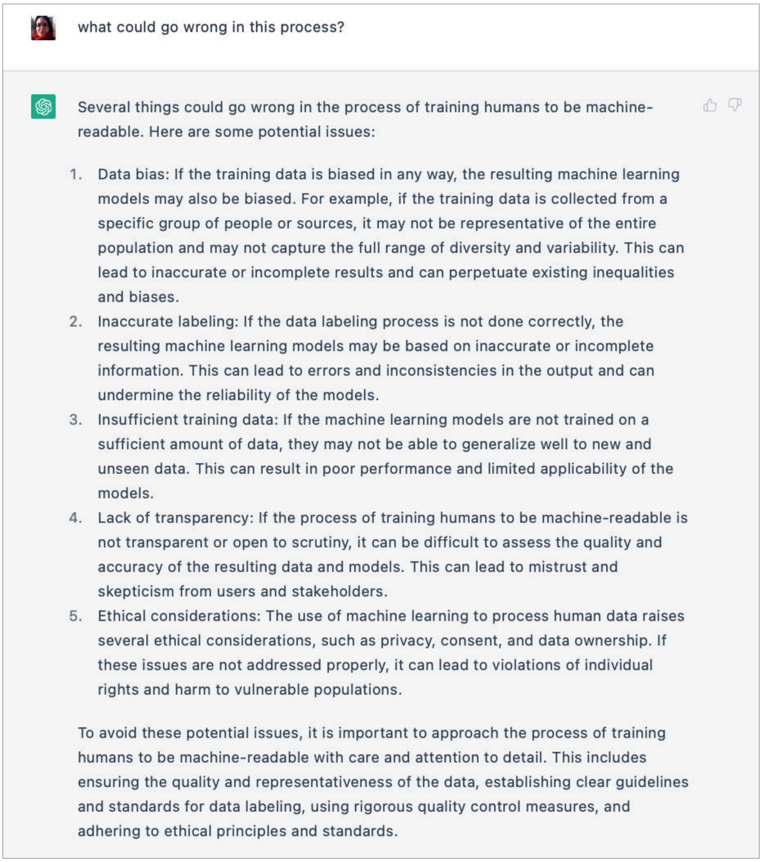
It is known that because we construct ourselves through social media, as image and circulation strategy, the ‘self’ is becoming dispersed, rhizomatic, frameable, so as to be machine readable (see Giannachi 2022). Crawford and Paglen’s *Training Humans* (2020), the first major photography exhibition devoted to the exploration of the collections of images that are used by scientists to train AI systems, is a clear step forward in acknowledging the kind of framing devices we create for AI to see and categorise the world, revealing a wealth of information about how AI systems perpetuate social classification and injustice, surveillance and control, with the risk of echoing phrenology and eugenics of the past. I checked with GPT-4 how it thought that humans could be trained to be machine readable, and the reply comprised advice on data collection, pre-processing, labelling, training and refinement, a combination of archival and R&D methodologies aimed at delivering **[fig. 8]** the sort of meta-archive I postulate in *Archive Everything* (2016).

GPT-4 suggests that to ensure humans can be trained to be properly machine readable it is necessary for the labelling to be accurate, something requiring significant quality control that at present is not part of the proposition. So, then the question arises as to who is responsible for the labelling and how the process ought to be framed to work for both machines and humans.

## 5 Training Humans for AI: Luca Viganò's *The First*

The Italian cybersecurity expert and playwright Luca Viganò's short film *The First* (2021) analyses how we frame our relationship with machines by turning the Turing test around. While the Turing test is about whether a machine can deceive a human into thinking that they too are human, the Inverse or Reverse Turing test is to see whether a human can persuade a machine that they are a machine. So, in the film the character called girl tries to persuade the machine that she is a machine so she can become a 'pilot', which her world forbids in that, as the voice off screen suggests, "pilots, doctors, surgeons, accountants, builders… not for us anymore… to protect us". So, the film portrays a future where humans are excluded from society for their protection and the protection of machines, raising the question as to who the 'us' that is being referred to actually is. Despite the use of binary colours and metaphors, it is no longer possible here to clearly distinguish between frames. While we may in fact think everything is binary (black/white, left/right, front/back) we soon understand this is not so simple. Hence there is also another voice, articulating the chronology towards an impact, which catapults us towards an end in which everything is erased though we are left wondering, if there was a first, whether there may be others.

The title of the film is *The First* but in fact there are two firsts. The girl is the *first* human to persuade a machine that she is a machine (she learns how machines behave so as to become one of them). The machine in the story is the *first* to make the mistake of accepting the girl as a machine. So, in the world portrayed by this film machines and humans may no longer be clearly distinguishable by us or even by themselves. Or maybe they are, but they choose to ignore that. At the heart of AI is the concept of deceit, which is built in the machine but is also connected to the innate desire humans have to locate intelligence in machines and to redefine, reframe, themselves in order to do so. But the world that is being designed is our own, and if I were to hold up a mirror to that world, I would see the self-destruction produced by viruses, wars, climate change that we produced.



what could go wrong in this process?

Several things could go wrong in the process of training humans to be machine-readable. Here are some potential issues:

1. Data bias: If the training data is biased in any way, the resulting machine learning models may also be biased. For example, if the training data is collected from a specific group of people or sources, it may not be representative of the entire population and may not capture the full range of diversity and variability. This can lead to inaccurate or incomplete results and can perpetuate existing inequalities and biases.
2. Inaccurate labeling: If the data labeling process is not done correctly, the resulting machine learning models may be based on inaccurate or incomplete information. This can lead to errors and inconsistencies in the output and can undermine the reliability of the models.
3. Insufficient training data: If the machine learning models are not trained on a sufficient amount of data, they may not be able to generalize well to new and unseen data. This can result in poor performance and limited applicability of the models.
4. Lack of transparency: If the process of training humans to be machine-readable is not transparent or open to scrutiny, it can be difficult to assess the quality and accuracy of the resulting data and models. This can lead to mistrust and skepticism from users and stakeholders.
5. Ethical considerations: The use of machine learning to process human data raises several ethical considerations, such as privacy, consent, and data ownership. If these issues are not addressed properly, it can lead to violations of individual rights and harm to vulnerable populations.

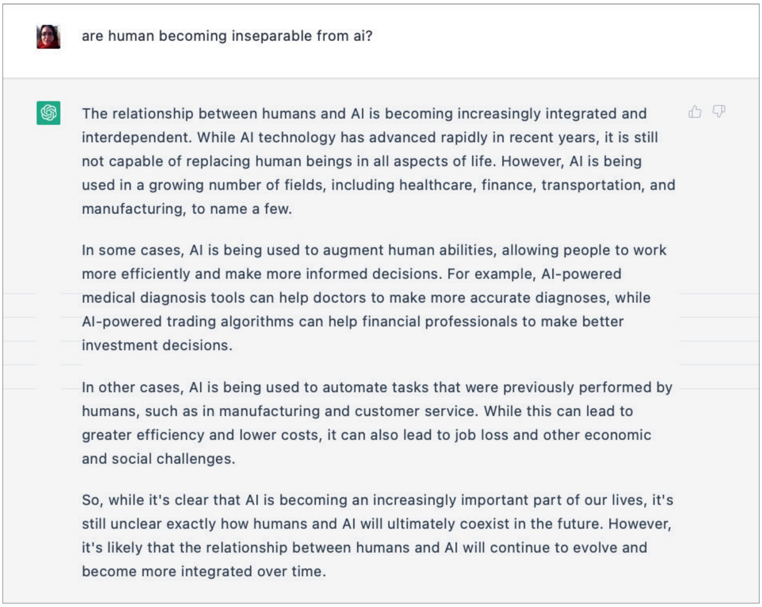
To avoid these potential issues, it is important to approach the process of training humans to be machine-readable with care and attention to detail. This includes ensuring the quality and representativeness of the data, establishing clear guidelines and standards for data labeling, using rigorous quality control measures, and adhering to ethical principles and standards.

**Figure 9** GPT-4, screenshot, 2023

So, the question is not only how we can frame ourselves though machines so that machines read the best and not the worst in us but also what might go wrong in this process. I ask GPT-4, and the answer is quite comprehensive.

So, training humans for AI, according to AI, can include several critical issues to do with data bias, inaccurate labelling, the insufficient training of data, lack of transparency and ethics, which tend to be defined by humans **[fig. 9]**. So, then, the question arises as to whether humans are not only being trained but also framed. This deliberately ambivalent question results, as GPT-4 points out, in several complex possibilities.

GPT-4's answer shows that humans are both the source and outcome of these AI-human entanglements in that they are the originating framework, its creators and users, and the perpetuators of



are human becoming inseparable from ai?

The relationship between humans and AI is becoming increasingly integrated and interdependent. While AI technology has advanced rapidly in recent years, it is still not capable of replacing human beings in all aspects of life. However, AI is being used in a growing number of fields, including healthcare, finance, transportation, and manufacturing, to name a few.

In some cases, AI is being used to augment human abilities, allowing people to work more efficiently and make more informed decisions. For example, AI-powered medical diagnosis tools can help doctors to make more accurate diagnoses, while AI-powered trading algorithms can help financial professionals to make better investment decisions.

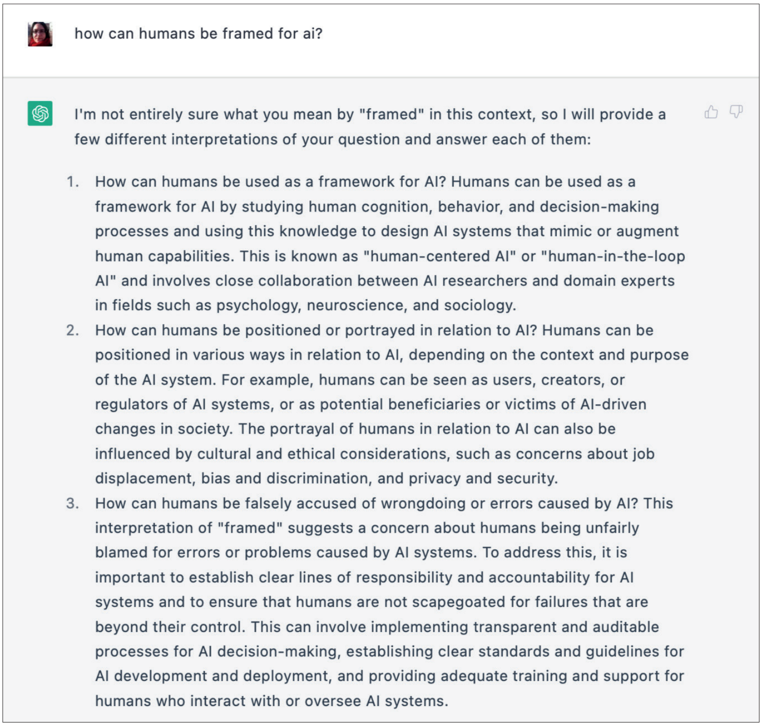
In other cases, AI is being used to automate tasks that were previously performed by humans, such as in manufacturing and customer service. While this can lead to greater efficiency and lower costs, it can also lead to job loss and other economic and social challenges.

So, while it's clear that AI is becoming an increasingly important part of our lives, it's still unclear exactly how humans and AI will ultimately coexist in the future. However, it's likely that the relationship between humans and AI will continue to evolve and become more integrated over time.

**Figure 10** GPT-4, screenshot, 2023

its potentially dubious ethical findings **[fig. 10]**. And just as we are invariably producing a climate emergency, we are also framing humans through AI, at the risk of perpetuating, within the world created by it, all our prejudices, discriminations, and errors. I ask GPT-4 one last question to do with as to whether this world already exists, or whether it is still a science fiction. From the answer **[fig. 11]** it is clear that the integration between humans and AI has already started and while GPT-4 points out that it is an evolving relation, it also notes that in some areas the integration is already fairly advanced. And while it highlights the benefits that may result from faster processing, it also points out that it can lead to job loss and "other economic and social challenges". And then again, in the last paragraph, it throws me by using the word *our* in relation to the predicament that AI is becoming increasingly part of 'our' lives. Does this mean that AI, as in Vigano's film, might one day think it is human? Or does it mean it knows, as the humanoid-robot AI-Da, created by the computer scientist Aidan Meller, that while it may be capable of drawing self-portraits by using a camera eye and a pencil, the real issue is what does it mean that it can do this since it has no self in the first place (Meller 2021).

As the philosopher Emanuele Coccia suggests, while in the 20th century the 'I' was "the place and medium" through which we could



how can humans be framed for ai?

I'm not entirely sure what you mean by "framed" in this context, so I will provide a few different interpretations of your question and answer each of them:

1. How can humans be used as a framework for AI? Humans can be used as a framework for AI by studying human cognition, behavior, and decision-making processes and using this knowledge to design AI systems that mimic or augment human capabilities. This is known as "human-centered AI" or "human-in-the-loop AI" and involves close collaboration between AI researchers and domain experts in fields such as psychology, neuroscience, and sociology.
2. How can humans be positioned or portrayed in relation to AI? Humans can be positioned in various ways in relation to AI, depending on the context and purpose of the AI system. For example, humans can be seen as users, creators, or regulators of AI systems, or as potential beneficiaries or victims of AI-driven changes in society. The portrayal of humans in relation to AI can also be influenced by cultural and ethical considerations, such as concerns about job displacement, bias and discrimination, and privacy and security.
3. How can humans be falsely accused of wrongdoing or errors caused by AI? This interpretation of "framed" suggests a concern about humans being unfairly blamed for errors or problems caused by AI systems. To address this, it is important to establish clear lines of responsibility and accountability for AI systems and to ensure that humans are not scapegoated for failures that are beyond their control. This can involve implementing transparent and auditable processes for AI decision-making, establishing clear standards and guidelines for AI development and deployment, and providing adequate training and support for humans who interact with or oversee AI systems.

**Figure 11** GPT-4, screenshot, 2023

experience the world (Coccia 2021, 84) now we are immersed in a kind of “collective novel in which everybody is author, character and reader of how their lives intertwine with those of others” (Coccia 2021, 84, Author’ transl.). This means that reality and fiction are no longer separable (Coccia 2021, 84), and consciousness is no longer associated with the ‘I’ but it is outside of us (Coccia 2021, 86) devolved to machines. The world, Coccia continues, is “no longer composed of events, but of a diffused psyche, of a consciousness in which we are all immersed” (Coccia 2021, 87). We are no longer entirely separable from the machines we have created and the AI that regulates and frames them **[fig. 10]**. We may have always already known those machines, which we may have created to defend us from the unknown. Or, as in the case of The Turk, we may have actually informed the operation of those machines. Or maybe even, as in the case of Capek’s robots, we may have let go of our ‘selves’ for those machines. In any case, there is no longer a frame between us and AI. The AI is *us*. Or, at least, it thinks it is…

## Bibliography

Bowden, M.A. (2006). *Mind As Machine: A History of Cognitive Science*. Oxford: Oxford University Press.

Bratton, B.H. (2016). *The Stack: On Software and Sovereignty*. Cambridge (MA): MIT Press.

Coccia, E. (2021). *Metamorfosi.* Trad. di S. Mambrini. Torino: Einaudi.

Colby, K.M. et al. (1972). “Turing-Like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes”. *Artificial Intelligence*, 3, 199-221.

Crawford, K.; Paglen, T. (2019). “Excavating AI: The Politics of Training Sets for Machine Learning”. `https://excavating.ai.`

Damasio, A. (1994). *Descartes’ Error*. New York: Avon Books.

Damasio, A. (2003). “Feelings of Emotion and the Self”. LeDoux, J.; Debiec, J.; Moss, H. (eds), *The Self: From Soul to Brain*. New York: The New York Academy of Sciences, 253-61.

Floridi, L. (2013). “Present and Foreseeable Future of Artificial Intelligence”. Sabba, F. (ed.) (2013), *Noetica VS Informatica = Atti del Convegno Internazionale* (Roma 19-20 novembre 2013). Firenze: Olshki, 131-6.

(Roma 19-20 novembre 2013). Firenze: Olshki, 131-6

Giannachi, G. (2016). *Archive Everything.* Cambridge (MA): The MIT Press.

Giannachi, G. (2022). *Technologies of the Self-Portrait,* London; New York: Routledge.

Goodfellow, I.J. et al. (2014) “Generative Adversarial Networks, 1-9”. *arXiv preprint arXiv:1406.2661*.

Han, B.-C. (2022). *Le non cose*. Torino: Einaudi.

Hershman Leeson, L. (2005). “Agent Ruby”. Tromble, M. (ed.), *The Art and Films of Lynn Hershman Leeson.* Berkeley: University of California Press, 92-4.

Hershman Leeson, L. (2014). *Robot Dialogue: Hershman and Agent Ruby*. `https://www.youtube.com/watch?v=XLOgMgCNC_w.`

Kaasinen, E. et al. (2022). “Smooth and Resilient Human-Machine Teamwork as an Industry 5.0 Design Challenge” *Sustainability* 14, 5(2773). `https://doi.org/10.3390/su14052773.`

Klingemann, M. (2019). *Circuit Training*. `https://vimeo.com/338883309`.

Lyytinen, K.; Nickerson, J.V.; King, J.L. (2021). “Metahuman Systems = Humans + Machines That Learn”. *Journal of Information Technology*, 36(4), 427-45. `https://doi.org/10.1177/0268396220915917`.

Manovich, L. (2018). *AI Aesthetics*. New York: Strelka Press.

Meller, A. (2021). *Who is Ai-Da?*. `https://www.ai-darobot.com/about`.

OpenAI Charter (2018). `https://openai.com/charter`.

OpenAI (s.d.). `https://openai.com/blog/planning-for-agi-and-beyond`.

OpenAI (2023). `https://arxiv.org/pdf/2303.08774.pdf`.

Paglen, T. (2019). *From ‘Apple’ to ‘Anomaly’ (Pictures and Labels). Selections From the ImageNet Dataset for Object Recognition*. London: Barbican.

Pinotti, A. (2017). “Self-Negating Images: Towards An-Iconology”. *Proceedings of the International and Interdisciplinary Conference*, 1(9). `https://doi.org/10.3390/proceedings1090856`.

Schopman, J. (1987). “Frames of Artificial Intelligence”. Bloomfield, B. (ed.), *The Question of Artificial Intelligence: Philosophical and Sociological Perspectives.* London: Croom Helm, 165-219.

Strecker, A. (2017). “An Urgent Look at How Artificial Intelligence Will See the World”. *LensCulture*. `https://www.lensculture.com/articles/trevor-paglen-an-urgent-look-at-how-artificialintelligence-will-see-the-world`.

Suchman, L. (2006). *Human-Machine Reconfigurations: Plans and Situated Actions.* 2nd ed. Cambridge: Cambridge University Press).

Tromble, M. (ed.) (2005). *The Art and Films of Lynn Hershman Leeson.* Berkeley: University of California Press.

Turkle, S. (1988). “Artificial Intelligence and Psychoanalysis: A New Alliance”. Graubard, S.R. (ed.), *The Artificial Intelligence Debate, False Starts, Real Foundations.* Cambridge (MA): The MIT Press, 241-68.

Weizenbaum, J. (1966). “ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine”. *Communications of the ACM*, 9(1), 36-45.

Zylinska, J. (2020). *AI Art, Machine Visions and Warped Dreams.* London: Open Humanities Press.