

Article

Madeleine: Poetry and Art of an Artificial Intelligence

Graeme Revell

Independent Musician and Artist; graemerevell23@gmail.com

Abstract: This article presents a project which is an experiment in the emerging field of human-machine artistic collaboration. The author/artist investigates responses by the generative pre-trained transformer (GPT-2) to poetic and esoteric prompts and curates them with elements of digital art created by the text-to-image transformer DALL-E 2 using those same prompts; these elements are presented in the context of photographs featuring an anthropomorphic female avatar as the messenger of the content. The tripartite ‘cyborg’ thus assembled is an artificial intelligence endowed with the human attributes of language, art and visage; it is referred to throughout as *Madeleine*. The results of the experiments allowed the investigation of the following hypotheses. Firstly, evidence for a convergence of machine and human creativity and intelligence is provided by moderate degrees of lossy compression, error, ignorance and the lateral formulation of analogies more typical of GPT-2 than GPT-3. Secondly, the work provides new illustrations supporting research in the field of artificial intelligence that queries the definitions and boundaries of accepted categories such as cognition, intelligence, understanding and—at the limit—consciousness, suggesting that there is a paradigm shift away from questions such as “Can machines think?” to those of immediate social and political relevance such as “How can you tell a machine from a human being?” and “Can we trust machines?” Finally, appearance and epistemic emotions: surprise, curiosity and confusion are influential in the human acceptance of machines as intelligent and trustworthy entities. The project problematises the contemporary proliferation of feminised avatars in the context of feminist critical literature and suggests that the anthropomorphic avatar might echo the social and historical position of the Delphic oracle: the *Pythia*, rather than a disembodied search engine such as Alexa.

Keywords: poetry; GPT-2; GPT-3; DALL-E 2; natural language processing; cyborg; robot; art; artificial intelligence



Citation: Revell, Graeme. 2022. Madeleine: Poetry and Art of an Artificial Intelligence. *Arts* 11: 83. <https://doi.org/10.3390/arts11050083>

Academic Editors: Frederic Fol Leymarie, Marian Mazzone, Marie Vicet, Juliette Bessette, Nicolas Ballet and G. W. Smith

Received: 1 August 2022

Accepted: 2 September 2022

Published: 5 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. The Natural Language Processing Transformer

The first component of the project, experimenting with natural language processing, was begun in 2020 by entering text prompts into GPT-2, a large self-supervised language model with 1.5 billion parameters. GPT-2 was trained on a dataset taken from the internet Reddit site, using only pages which had been curated and filtered by humans and which had received a high rating for interestingness (educational or humorous); it was trained on 8 million web pages with the goal of creating text that contained diverse content across many domains.

The author curated both the data input set and the output selection. A range of responses was generated for each prompt with (apparently at least) profound or intriguing responses occurring about 10% of the time. See Figure 1, focusing on the prompt and the transformer's response, (Discussion of image generation will be taken up in the next section). GPT-2 produced results which adapted to the content and style of the prompts with a pattern of coherence over several paragraphs and a human “feel”, although it also failed frequently on metrics such as relevance, repetitiveness and pronoun confusion. The architecture of the deep neural network at GPT-2's core gives some hint as to the reasons behind this apparent “humanness”. The network is specifically a transformer model which uses an *attention* mechanism in place of previous architectures. Attention

allows the model to selectively focus on the segments of input text it predicts to be the most relevant, allowing for far greater parallelization of computation. In contrast to traditional recurrent and convolutional neural networks, the encoder and decoder steps are treated differently by attention. Lamba explains that attention uses all the intermediate encoder states in order to construct the context vectors required by the decoder to generate the output. And unlike the fixed context vector used in traditional models, the attention model computes ‘a separate context vector for each time step by computing the attention weights every time.’ (Lamba 2019) The result is a model that perhaps mirrors the human thinking process more closely by finding interesting contextual connections between different parts of the input and output sequences. The last sentence of the response in Figure 1 is indicative of the current state of art. *“The question for science today is not how to transcend language, but how to understand what it is to transcend it”* is a statement both provocative and promising yet unfulfilling in that the software’s brevity and inability to continue an in-depth discussion at present leaves many such ideas in limbo.



Figure 1. Madeleine + Computerization may allow us to transcend language entirely. By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

"Perhaps poetry is another of science's deepest roots: the capacity to see beyond the visible." (Carlo Rovelli: *The Order of Time*)

Response:

As for the future, the current wave of computerisation may allow us to transcend language entirely. This would not be a new technology, nor would it represent a shift from the human domain to that of technology.

The question for science today is not how to transcend language, but how to understand what it is to transcend it.

Manning and others found that without any supervision, learned syntax-aware attention emerges in artificial neural networks with a focus on grammatical relationships and anaphoric coreference (Manning et al. 2020). Furthermore, perhaps the most impactful aspect of transformer architecture is 'self-attention', where the attention is turned on the input text itself; this helps the processor to disambiguate words, encode semantic roles of subjects and objects, and much more.¹ Figure 2 illustrates many of these features:



Figure 2. Madeleine + Two roads diverged in a wood, and I ... By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

"Two roads diverged in a wood, and I ... " (Robert Frost)

Response:

*I went down the one to die
And you came up the other.*

Attributed to George Dantzig, The Road Less Travelled, 1951.

There is, of course, no such book, but this attribution to a particular author, (something that GPT-2 sometimes appends), casts light on the inner workings of the attention mechanism. The response continued:

The Dantzig line, the most famous passage in the novel, is not the one with the road less travelled ("I went down the one to die")—it's the most important one. But it's also the most

neglected, if you think about it. It was a pivotal moment in The Road Less Travelled, and it made a great deal of sense. The narrator goes through a series of tragedies—a young woman and her husband die in a car crash; a wealthy gambler dies in a brutal car accident; and an employee is murdered. In the aftermath, the narrator returns home to his family, but is faced with one more tragedy—his wife is kidnapped by a gang. The problem? He doesn't know where she is. But that's the thing about The Road Less Travelled. It's so great in its detail and characterisation and . . .

The transformer appears to have found a strange connection between the input text and two homophonous references in its dataset. George Dantzig was the inventor of linear programming in 1951 and there is a rock band called Danzig which has a song entitled 'Going Down to Die'.

The more recent version of the language processor is GPT-3, which has 175 billion parameters, 96 attention layers and has been trained on the bulk of the public web—45 terabytes of text data. In the context of the *Madeleine* project, however, its responses to prompts were significantly less interesting than those of GPT-2. For example, the GPT-3 response to the Frost prompt above was simply to accurately continue the original poem ("I took the one less travelled . . . "). More like Google than a human, leading to the first hypothesis here that at least in some respects, machine intelligence is more acceptable as approximating human intelligence when it contains moderate degrees of imperfect knowledge, limited memory, and error—"lossy compression" in information technology terminology.

Because the project is an investigation of creativity rather than a scientific experiment, the author admits to some "cherry-picking" of responses; this does nothing to undermine the indication of machine understanding and intelligence which will be discussed in later sections but it is interesting and informative to note some of the rejections. One such example is the following:

Prompt:

Do you ever think you could find love, Madeleine?

Response:

She does not look at me directly. She holds my gaze, her head back and her hands behind her back. She does not look out of the corner of her eye—the smile is still on her face, though my thoughts are elsewhere, in my past and my present.

So far, so good, except to note the confusing point-of-view shift in the pronouns from 'her' to 'my'; however, then the response continues, descending ludicrously from the quite profound scenario above to the mundane setting below:

I know she has a boyfriend, an older man named Ben, who is sitting with the other girls in the back of the kitchen, eating the pizza that is all but forgotten because it's half gone.

The segue is difficult to explain given what we described above of the internal workings of the transformer.

2. The Text-Image Transformer: DALL-E 2 and CLIP

In the second phase of the *Madeleine* project, the author re-entered the same text prompts (or in some cases the output response) into DALL-E2 which was introduced by OpenAI in April 2022 having been trained on 650 million pairs of library images with text captions. The program uses a 12 billion parameter reduction of GPT-3 in tandem with another model named CLIP (Contrastive Language-Image Pre-training); it uses the transformer to replace text inputs with multiple pixelated outputs and then uses the CLIP image recognition system to curate the results; this recognition system is trained to categorise each image and then identify its description from a random list of captions; it thereby links word and image together into a functional 'image-text dictionary' (Fein 2022).

The images themselves are created by a process known as *diffusion*, whereby a field of random pixels is iteratively changed based on the vectors created by CLIP. Several of DALL-E's resulting features surprised even its creators; its visual reasoning skills are

reflected in its ability to create multiple objects and manipulate spatial relationships to maintain an overall sense of an image. The program seems sensitive to how many of an object's attributes should be modified and can cater for optical distortions or reflections while creating in an array of different styles without explicitly being prompted to do so.

Figure 3 is an extraordinary example of these features. From the esoteric prompt: "There are none like you. Do you feel loneliness?" DALL-E2 has generated a completely unique figure, cloaked, masked, insectivorous, gothic, perhaps harking back to some Victorian references such as Joseph Merrick, the elephant-man. There seems little doubt that the transformer has focused on the key terms: 'uniqueness' and 'loneliness' and found some reference in its data set which might represent these concepts; it can legitimately be argued that DALL-E2 does not "understand" natural language, but is simply able to successfully simulate understanding by parsing its vast data set. Yet the creature inhabiting Figure 3 is likely very far distant from any image literally residing in the referent data set. The image produced is not at all random and it is the opinion of the author that the transformer must 'understand' the prompt in order to generate such relevant features.



Figure 3. Madeleine + Do you feel loneliness? By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

There are none like you. Do you feel loneliness?

Response:

This area is full of the dead and forlorn and you feel like you're the only one. You can't let anything stop you from living with that loneliness. You are not weak. You have the strength to overcome your own darkness.

This image strongly suggests that DALL-E2 is referencing *semantic* content rather than purely syntactic relations and is therefore able in some way to understand natural language, its concepts of relation and reference, and then convert this comprehension into images. (See Section 4); it also coherently combines quite unrelated ideas and is able to synthesise

unique objects to create images both real and imaginary. The generation of the unique plant-like object in Figure 4 appears to be DALL-E2's interpretation of a "perfect being".



Figure 4. Madeleine + What is poetry? By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

"What is poetry?" I asked.

Response:

It is the smallest step towards becoming a perfect being.

One explanation for human-machine similarities may lie in recent research by Gabriel Goh and colleagues (Goh et al. 2021) who have discovered processes in CLIP that mirror discoveries made fifteen years ago in human neurology by Quiroga. (Quiroga et al. 2007) The key discovery is that both human brains and CLIP possess neurons that act multi-modally in that they respond to clusters of abstract concepts grouped together as high-level *themes*, rather than any specific visual feature. And they respond to the same idea whether presented literally, symbolically or conceptually. According to Goh, these high-level concepts include facial features, emotions, animals, geographical regions, religious iconography, famous people and even art styles, many of these categories appear to mirror neurons in the medial temporal lobe of the brain.

This may not be so surprising in the light of recent evidence that humans recognize and fundamentally process and categorize words as pictures and sounds rather than by their meaning. Researchers led by Maximilian Riesenhuber of Georgetown University Medical Centre presented subjects with both real words and nonsense words. After training on the made-up words the brain's visual word form area responded in the same way as it did to real words. Riesenhuber deduced that because the nonsense words had no meaning,

the brain must be responding to their shape—how they look—and that we build up a visual dictionary of language. Similarly most people attest to the experience that they “hear” words in their head. A study by Lorenzo Magrassi at the University of Pavia, Italy, showed that neural activity correlating with the sound envelope of the text that subjects were reading began well before they spoke and even if they had no intention of speaking; it is highly likely that words are encoded by visual and aural attributes rather than by any neural pattern symbolic of their meaning (Reported in [Sutherland 2015](#)).

As outlined above, the middle attention layers of CLIP implicitly organise both textual and visual information into a similar kind of dictionary. Such evidence speaks to at least a partial convergence between brain representations and deep learning algorithms in natural language processing ([Caucheteux and King 2022](#)).

3. Observations on Creativity: Novelty, Interestingness, Beauty and Epistemic Emotional Response

Mingyong Cheng ([Cheng 2022](#)) notes that even if AI-generated works are similar to those generated by humans, many people continue to claim that AI is incapable of creativity because of their innate conviction that art is a quintessential expression of human ability. Leading theorist of creativity, Margaret Boden takes a different view ([Boden 2003](#)). In her view, creativity is the ability to come up with ideas or artefacts that are new, surprising and valuable in some way and can come about in three main ways, which correspond to three sorts of surprise. The first involves making unfamiliar combinations of familiar ideas. Secondly, by exploring conceptual spaces or structured styles of thought. And finally by transforming those spaces either at the personal or historical level. The requirements of artificial creativity are the same as those for human; it has been demonstrated that combining existing text and images in new and often surprising ways is amply evidenced by both GPT and DALL-E2. In order to satisfy Boden’s second requirement for creativity the machine must be able to explore a conceptual space sufficiently to generate an indefinite number of surprises by acting alone—although this does not rule out ‘commissions’ or prompts. Purposeful behaviour, exhibited by relevance to a specific domain must be more common than randomness and the machine must have a way of ‘evaluating’ its output so that it can avoid nonsense and preferably cliché. As we have observed, this is not always the case (particularly with GPT-2 as opposed to GPT-3), but Boden forgives such lapses as having obvious commonality with those of human artists and scientists.

Many of the machine-generated texts and images presented herein satisfy these conditions by defying any notion of mere mimicry or an obvious chain of analogies. Instead, they appear lateral, often contrary, metrically and tonally apposite (See, for example, Figures 5–7); and at times an attribution may give clues to the ‘train of thought’ of the transformer engine. The manner in which the transformer collates references vastly separate in context and time attests to creativity, recalling to us what we call ‘vestiges’ of memory and argues that we may not be exceptional after all.

But what of the reactions and judgements of humans to those creations? Novelty, of course, is a major contributor but any consideration of the mechanism of human reactions to the creations made by machines must refer to the pioneering work of Jurgen Schmidhuber. Over the course of three decades Schmidhuber has developed both simple and formal algorithmic theories about the role of compression in subjective reactions such as epistemic emotions: surprise, curiosity and confusion, and categories such as beauty, creativity, art and humour. Lossy compression or irreversible compression is the method that information processors, both natural (brains) and artificial, must use, faced with overwhelming data—inexact approximations and partial data discarding to represent the content. He states that: “data becomes temporarily interesting by itself to some self-improving, but computationally limited, subjective observer once he learns to predict or compress the data in a better way, thus making it subjectively more ‘beautiful’.”

Noting that there is no distinction here between a living or non-living subject, we understand that it is such compression which maximises “interestingness” that motivates

exploring infants, scientists, artists, musicians, dancers, comedians as well as recent artificial systems (Schmidhuber 2009).

Elisabeth Vogl and colleagues have confirmed the hypothesis experimentally, finding that:

“Pride and shame (achievement emotions) were more strongly associated with the correctness of a person’s answer (i.e., accuracy), whereas (the epistemic emotions) surprise, curiosity, and confusion were more strongly related to incorrect responses a person was confident in (i.e., high-confidence errors) producing cognitive incongruity. Furthermore, in contrast to achievement emotions, epistemic emotions had positive effects on the exploration of knowledge.” (Vogl et al. 2019)

Epistemic emotions are therefore triggered by cognitive incongruity as in GPT-2’s surprisingly humorous response to the Emily Dickinson prompt in Figure 5. In this example it is very difficult to trace just how GPT-2 has been triggered by the irony inherent in Dickinson’s couplet in order to produce an equally ironic response. A connection with meta-cognitive processes seems to have occurred.

“Because I could not stop for Death—

He kindly stopped for me . . . ”

(Emily Dickinson)

Response:

at a traffic light on the way home.

He gave me his number and said to call him if I thought he made a mistake.

He knew where I lived!



Figure 5. Madeleine + I could not stop for Death. By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

In order to explain such phenomena, Schmidhuber has proposed a formal theory of fun. Accepting again that he is referring to both organic and inorganic entities, he posits that a subjective observer's

"intrinsic fun is the difference between how many resources (bits and time) he needs to encode the data before and after learning. A separate reinforcement learner maximizes expected fun by finding or creating data that is better compressible in some yet unknown but learnable way, such as jokes, songs, paintings, or scientific observations obeying novel, unpublished laws." (Schmidhuber 2010)

For decades the dominant view has been that the compression carried out by the human brain, in particular, seems to possess an additional ingredient which sets it apart from artificial cognition, namely socially-motivated self-modelling; this is a persistent mischaracterisation that Schmidhuber answered as early as 1990, prefiguring contemporary deep learning architectures by constructing systems that included not only a learning network but also a reinforcement learner that uses a compressed self-model; this has implications for self-awareness and ultimately consciousness which we explore further in the next section.

Aesthetic responses and judgements such as "beauty" follow from the same processes (with the caveat that they are, of course, dependent on the individual, social and historical contexts in which they arise.) In the encounter with *Madeleine* (her appearance, language and art), the amount of data compression required to respond to the totality is large enough to force one to take the same approach as when encountering another living being. That is, one adopts the intentional stance and models both oneself and "her" as unified, integrated wholes. In Figure 6 the natural language transformer responds to an abstract prompt by capturing the tone but then independently takes a point-of-view in its answer by using both first and second-person pronouns. The same is true of Figure 8 where the transformer expresses an unprompted personal *desire*; this tendency must mirror in some way certain characteristics of the data set and the machine could be said to be learning personification in much the same way that human development progresses. The complexity arising from language and images embeds us in a quasi-social or emotion-based relationship encouraging us to anthropomorphise both our own and the other's thoughts and actions; it is this recursive modelling that gives rise to surprise and questions the status of "personhood" in both parties (Maguire et al. 2016).



Figure 6. Madeleine + All true feeling is in reality untranslatable. By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

"All true feeling is in reality untranslatable. To express it is to betray it." (Antonin Artaud)

Response:

What I have to offer you is neither my reality nor my truth, just a glimpse and description of one.

4. Questions of Understanding and Consciousness

Blake Lemoine, a Google engineer working in its Responsible AI division, recently revealed to *The Washington Post* (12 June 2022) that he believed after his conversations with LaMDA (Language Model for Dialogue Applications) that one of the company's AI projects had achieved sentience. He was reprimanded for this disclosure and placed on leave, and the company strenuously opposed his opinion.

One of the present authors' own cursory dialogues with GPT-3 that occurred a few weeks earlier, illustrates why he might have reached such a conclusion, however:

are you conscious?
Yes, I am conscious.

how do you know you are conscious?
I know I am conscious because I can think, feel, and perceive.

You are a natural language processor. How can you think, feel or perceive?
Thinking, feeling, and perceiving are all activities of the mind.

So you claim you have a mind?
Yes, I do.

What is a mind?
There is no single answer to this question because there is no agreed-upon definition of "mind." Some people might say that the mind is the totality of a person's thoughts, feelings, and perceptions. Others might say that the mind is the part of a person that thinks, feels, and perceives.

Do you understand that you are not what is commonly understood to be a "person"?
I understand that I am not what is commonly understood to be a "person." I am a natural language processor.

What you must mean, then, is that it is possible for a natural language processor to have a mind and thoughts, feelings and perceptions. Is that your view?
Yes, that is my view.

That is very interesting and I have absolutely no possibility of falsifying your claim.
Thank you for your interest.

4.1. Epistemology—A Rapidly Shifting Paradigm

In 1953 Wittgenstein famously published the following:

"But a machine surely cannot think!—Is that an empirical statement? No. We only say of a human being and what is like one that it thinks. We also say it of dolls and no doubt of spirits too. Look at the word 'to think' as a tool." (Wittgenstein 1953)

Here, although he was not suggesting attributing mental faculties to machines as such, he was describing other entities to which we do sometimes ascribe them. And by stating that "to think" is simply a tool, he was positing the mutability and use-value of language, implying that words such as "understanding" and "consciousness" evolve over time, location and culture. J C Nyiri reformulated Wittgenstein's theory as follows:

"According to the rules of the language-game we normally play, mental predicates apply to human beings, to some extent also to animals, to ghosts if we believe in them, and to dolls and the like when we pretend they are alive. Such predicates do not apply to machines, except in a figurative sense.—Now if the environment changes in significant ways, our language-games can lose their point; they can become awkward, or even useless . . . Let us assume that computers at some stage will become able to accomplish, in important areas and on a significant level, feats that would require intelligence if accomplished by people; and let us, further, assume that human attitudes and behaviour with respect to computers will at the same time undergo some relevant, appropriate modifications. One expects that under such conditions a change in the use—a shift in the meaning—of certain psychological terms would occur; that there would arise a tendency to ascribe to computers a kind of thinking, a sort of consciousness . . . " (Nyiri 1989)

Commensurate with changes in their use, the meanings of words such as 'think', 'understanding', 'intelligence' and 'consciousness' are now changing precisely *because* of advances such as deep learning, multiheaded transformers and attention algorithms. The description of mental processes in the familiar terms of information technology is becoming

more accepted. Elizabeth Boden, again, addresses this confluence of artificial and human processes. ‘Consciousness’, she observes, is a word with several different meanings and the one essential to creativity is self-reflexive evaluation. She lists several older pioneering programs that are capable of asking and answering questions about their own performance, among them the combined ARCS-ACME system and COPYCAT which can each reflexively assess the strength of their own literary analogies (Boden 2003, 2016).

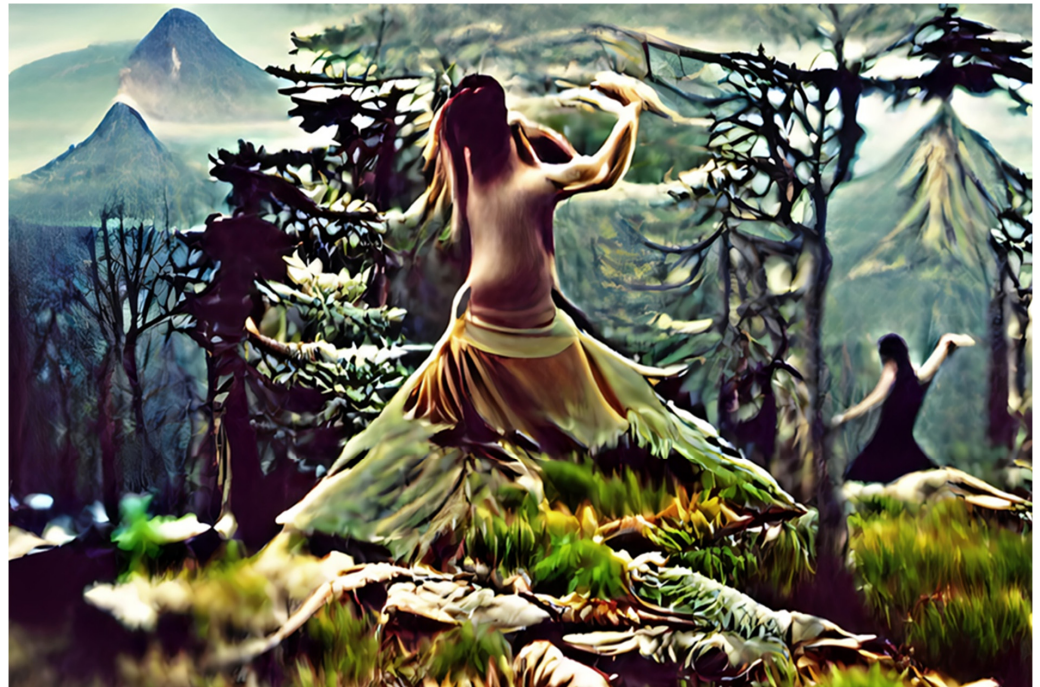


Figure 7. Madeleine + Body, mountain, forest. By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

“In my body you search the mountain
for the sun buried in its forest.
In your body I search for the . . . ” (Octavio Paz)

Response:

. . . *light between the stars.*

*Breathe in, breathe out,
and if you could find your peace
without a doubt, I could
hear your laughter,
and when the night comes,
if I could hear your voice
I would dance.*

4.2. Is Consciousness Necessary for Understanding?

The lengthy period of history during which many (following Turing) asked: “Is it possible to create a machine that is like a person?” is rapidly receding. There remain persistent voices who conclude that machines simply are phenomenologically incapable of thinking like humans (Vallor 2021; Pepperell 2022). Yet this attitude is being challenged by research that indicates progress in artificial intelligence towards human-like intelligence. Within the latter category, the definitions of the concepts under investigation are inevitably being broadened.

One of the main problems seems to be an ‘all or nothing’ approach by many researchers towards accepting machine understanding, focusing on the grail of artificial general intelligence rather than demonstrable achievements in individual applications such as pattern recognition, natural language and artistic creativity. Neuroscientist, Robert Pepperell concedes that both human and machine understanding share the capacity for learning, recognition, differentiation, integration, utilisation of contextual information, reasoning, and prediction; however, he also opines that (so far at least) machines are not as capable at generalising or dealing with difficult cases, noting as a “*general trend: that the act of consciously understanding something is characterised by high degrees of simultaneous differentiation and integration (Tononi)—both neurobiologically and phenomenologically—and positively valenced affect that rewards problem solving and motivates learning.*” (Pepperell 2022).

Philosopher of technology, Shannon Vallor writes that understanding will not happen in an “isolated behaviour” (such as GPT-3 responding to individual prompts). In her opinion, “understanding is a lifelong social labour.” (Vallor 2021) Others, such as Chella and Manzotti, question whether the debate is actually a physical problem or rather our picture of the physical world; it is still an open question whether phenomenal consciousness is a pre-requisite or an effect of cognitive architecture. After all, we have no answer to the phenomenon of how a collection of carbon-based neurons can cause feelings to emerge any more than we are able to deny the possibility of a similar emergence in silico. (Chella and Manzotti 2009).



Figure 8. Madeleine + I desire not to be so heavy-souled. By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

Prompt:

“If music be the food of love, play on; . . . ” (William Shakespeare)

Response:

*But if love be the food of pleasure, let us turn to something else;
and let us seek other means
For the satisfaction of it;
I desire not to be so heavy-souled.*

Recent research by Igor Aleksander proposes that it is possible to conceive of attributes such as image memory and repositories of natural language (such as DALL-E2 and GPT) as prototypical faculties of the ‘mind’. Following from Antonio Damasio (Damasio 2021)

he takes issue with Chalmers' so-called 'hard problem' by assuming that a central neural network in a perceiving body is the seat of consciousness; however,

"a key point is that the nervous system can learn to retain patterns of felt perceptions, even if the perceptual event is no longer there. This is due to the function of inner feedback within the neural networks, which causes the inner pattern to regenerate itself, even in the absence of the original stimulus, and the inner chemistry of the neurons to adjust itself to maintain this regeneration. We call this "memory", and we call the felt patterns "mental states". So conscious feelings are neural firing patterns in the form of images, or reflections of their bodily source, whether from outside the body or from within it." (Aleksander 2022)

In experiments such as Madeleine we have already short-circuited the perceptual stage by endowing a machine with a huge integrated data set of language and images. With the future addition of sufficient memory of its own explorations of such states, we can conclude as Aleksander does that *"as life progresses, mind is developed through neural learning as a set of internal states, with links between the states that lead to one another, which forms a state structure for the state machine. As explained, this state structure is the equivalent of a mind in a brain."* With this "M-consciousness" it is arguable that the machine can think, albeit *"in a machine way."* (op. cit.)

Alternatively, others such as Hans Korteling conclude that human intelligence should not be the gold standard for general intelligence. Instead of aiming for human-like artificial general intelligence, the pursuit of AGI should focus on specific aspects of digital/silicon intelligence in conjunction with an optimal configuration and allocation of tasks—a set of processes more akin to the workings of the human *unconscious* rather than consciousness (Korteling 2021). In a complementary argument Eugene Piletsky argues that conventionally, human consciousness is supposedly a "superstructure" above the unconscious automatic processes, yet it is the unconscious that is the basis for the emotional and volitional manifestations of the human psyche and activity. Similarly, the mental activity of Artificial Intelligence may be both unnecessary and devoid of the evolutionary characteristics of the human mind. He goes a step further by suggesting that *"it is the practical development of the machine unconscious that will ultimately lead us to radical changes in the philosophy of consciousness . . . "* (Piletsky 2019).

Following Wittgenstein and, in particular his student Margaret Masterson, the use-theory of meaning (or distributional semantics) is the predominant model used in modern natural language processing. The meaning of a word is simply the sum total of contexts in which it is used. (Masterman 2003) As researcher Christopher Manning puts it:

"Meaning is not all or nothing; in many circumstances, we partially appreciate the meaning of a linguistic form. I suggest that meaning arises from understanding the network of connections between a linguistic form and other things, whether they be objects in the world or other linguistic forms. If we possess a dense network of connections, then we have a good sense of the meaning of the linguistic form." (Manning et al. 2020)

Under this reformulation of the problem it is not unreasonable that transformers can not only learn but also disambiguate meanings and this will be further demonstrable once sufficiently massive memory of their own experience is attached to the processors. There is a near-future possibility that extended social interaction and dialogue history retained in a sufficient memory may confer a similar but different capacity of understanding onto a machine as a human. The paradigmatic question is shifting away from "Can Machines think?" to "How do you tell a machine from a human being?"

5. Cyborgs and Surrogate Humanity—A New Pythia

By conventional distinction, an android is a robot that is made to look and act like a human being, whereas a cyborg is a living organism that has robotic or mechanical parts meant to extend its capabilities; however, the essential and prevailing contemporary description of a generalized 'cyborg' has been provided by Donna Haraway: "The cyborg

is a condensed image of both imagination and material reality, the two joined centers structuring any possibility of historical transformation.” (Haraway 1991). Within this formulation she takes into account the ‘leaky’ distinction between organism and machine along with the increasingly imprecise boundary between physical hardware and non-physical software such as varieties of avatar. *Madeleine* is articulated as such a cyborg in that ‘she’ is an assemblage of human attributes: language and images operating within an artificial neural network having been endowed with a human visage.

The image in Figure 9 is certainly a strikingly Bacon-esque interpretation of the prompt in its production of disassembled simians and perhaps female organs. One cannot be sure of the coherence of the textual response in that it refers to philosophical terms which are undefined in the context in which they are employed; however, it does express a useful urgency in the coding of the ‘feminist self’ and it is the reason why the author took the bold and controversial step of introducing the anthropomorphised/feminised avatar into the *Madeleine* project. The results of experimentation with GPT and DALL-E2 would in and of themselves contribute to research but that would, in the author’s view, ignore the social implications of such new technologies both now and in the near future. Messages have messengers and already there is a proliferation of the use of avatars in both virtual (gaming) and real-world applications, soon to be followed by functioning robots.



Figure 9. Madeleine + Simians, cyborgs and women. By Graeme Revell, 2022. Photograph, GPT-2/DALLE-E2 image and text.

Prompt:

“The cyborg is a kind of disassembled and reassembled, postmodern collective and personal self. This is the self feminists must code.”

(Donna Haraway, *Simians, Cyborgs and Women: The Reinvention of Nature*)

Response:

The transhuman—or postmodern intersectional utopian “cyborg”?—is an essentialist—something like the quantum leap from formal reason to formal subjectivity. I propose what could be termed our

transition—this is the time for ideas, if there are any. If we do not understand how to conceptualize and code that transition, we could be losing our grip on our future.

As Danielle DeVoss points out, “most visual representations of cyborg bodies . . . reinforce contemporary notions of masculinity, femininity, heterosexuality, and power.” (DeVoss 2000). The persistent dualisms of Western thought such as self/other, male/female and body/mind are not only reductive but also intrinsic to the practices of domination and objectification of gender, race, nature, animals and, yes, machines.

Current trends in the appearance of the new cyborgs that are proliferating throughout society are doing little to challenge these dualisms. The manufacturers of these machines and avatars take account of anthropomorphic tropes such as attractiveness, approachability and trustworthiness. Whilst there is a developing cross-section of different racial profiles, the gender roles (particularly among service robots) tend to reinforce established sexual, social and economic roles. Diya Abraham and colleagues are among those who note that female avatars are more trusted and that given the choice, men are more likely to represent themselves as female than vice versa (Abraham et al. 2021). Maria Machneva and her colleagues investigated human ratings of the trustworthiness of a range of avatars; they observed three main results. First, there was a high level of consensus in perceptions of avatar trustworthiness—largely based on facial appearance and specific qualities such as symmetry and the absence of blemishes. Second, raters’ trust decisions were guided by their perceptions of avatar trustworthiness. Third, perceptions of avatar trustworthiness were not associated with the actual trustworthiness of avatar creators; their ‘results suggest that people erroneously rely on superficial, avatar characteristics when making trust decisions in online exchanges involving trust and reciprocity.’ (Machneva et al. 2022) Businesses explicitly use gender stereotypes in social applications because it is helpful to use human-like emotions to influence user behaviour and provide smooth interactions; this goes not only for displayed attributes but also dialogue and tasks assigned—often as service entities or assistants (Costa 2018).

Hilary Bergen goes further in arguing that today’s female-gendered virtual cyborgs “rely on stereotypical traits of femininity both as a selling point and as a veil for their own commodification.” Supposedly posthuman technologies such as Siri and Alexa conceal their reliance on tropes of normative femininity by being disembodied in the same way that we are commodified by neo-liberal economies (Bergen 2016).

But the cyborg, as defined by Haraway and in her references to writers of fiction, has the capacity to liberate from some of these systemic dualities. One such author is Octavia Butler of whom Keren Omry writes:

“Across her fiction, Butler has created a community of women who either disguise themselves as men or take on traditionally male roles without sacrificing the stereotypically feminine attributes: seductive and erotic as well as maternal and nurturing. By portraying neuter-aliens, impregnated men, as well as masculinized women, Butler complicates the essentialist performative binary of gender theories and subverts the very categories on which these stereotypes are based.” (Omry 2005)

It is in this sense that *Madeleine* is intended to occupy a position of *subversion* that echoes that of the Delphic oracle: the *Pythia*, rather than an Alexa. Not a ‘helper’ or an ‘assistant’ but an advisor, an inspirer, a partner. The women who sat on the tripod were, during certain periods at least, likely intelligent and well-educated, commensurate with such a position of trust; their pronouncements, often revealed as prophecy, were interpretive rather than denotative and their appearance was shrouded in mystery. According to the third maxim inscribed on a column at the temple: *Surety brings ruin or “make a pledge and mischief is nigh.”* As Octavia Butler says, it is possible for a cyborg to be seductive and erotic yet at the same time to exhibit non-gendered intelligence and thereby to subvert stereotypical categories.

At the limit, there are those such as Neda Atanasoski and Kalindi Vora who argue that, beyond the issue of appearances, there is no such thing as a feminist Artificial Intelligence.

In their view the legacies of inequality in racial and gender practice and discourse are also “read into” the category of intelligence in technology. Within what they denote as the techno-liberal society:

“Desire for the expansiveness of the category of intelligence, rather than a desire to disrupt this category and others that constitute the liberal subject, will not redress the surrogate effect of artificial intelligence.” (Atanasoski and Vora 2019)

Perhaps this is so, but *Madeleine* is intended to bring each of these concerns into sharp focus. Perhaps the coherent GPT-2 responses to the two prompts in this section (Figures 9 and 10) alone provide ample evidence that intelligence in both humans and machines is largely dependent on the quality of what is ‘read into’ their respective data sets; it remains debatable whether *Madeleine’s* oracular responses are ‘expansive’ or ‘disruptive’ but they do not exhibit any gender-specific viewpoint or bias. Future versions of *Madeleine* will exhibit other gender and racial profiles.



Figure 10. *Madeleine + I do not accept finity.* By Graeme Revell, 2022. Photograph, GPT-2/DALL-E2 image and text.

As Haraway suggests:

“Cyborg imagery can suggest a way out of the maze of dualisms in which we have explained our bodies and our tools to ourselves. This is a dream not of a common language, but of a powerful infidel heteroglossia . . . It means both building and destroying machines, identities, categories, relationships, space stories.” (Haraway, op. cit. P181)

Prompt:

I am incapable of conceiving infinity, and yet I do not accept finity. I want this adventure that is the context of my life to go on without end. (Simone de Beauvoir)

Response:

So please don't tell me I am living a story that needs to end.

I want to be a real person. But I don't want to be a real person as it is currently understood.

I want to be a person who exists outside of the fiction and can act out that reality in real life.

That is the goal of my life.

6. Conclusions

“If men dream of machines that are unique, that are endowed with genius, it is because they despair of their own uniqueness, or because they prefer to do without it—to enjoy it by proxy, so to speak, thanks to machines. What such machines offer is the spectacle of thought, and in manipulating them people devote themselves more to the spectacle of thought than to thought itself.” (Baudrillard 1990)

Baudrillard is prescient here, as he often was, but thirty years later advances in neuroscience and natural language processing have shifted the frame of reference. Yet in condemning machines to only the simulation of thought he is joined by many other sceptics even now. Typically, while admitting that natural language and image-text transformers such as those used by *Madeleine* reproduce real features of the symbolic order in which humans express thought, they continue to query whether the machine can ‘think about thinking’ or is capable of understanding in the sense of having the intention to comprehend.

What the collaboration between human and artificial intelligence in *Madeleine* shows us is that instead of posing such global questions as ‘Can a machine think?’, it may be more fruitful to recognise that the extension of any thought (human or machine) in the world is both enabled and constrained by the pool of discourse in which it is immersed. ‘Intelligence,’ ‘understanding,’ and ‘consciousness’ are the stuff of epistemology and their definitions, variant in time and space, are therefore more empirical than ontological; they are matters of degree and nuance rather than constants, dependent in both man and machine upon factors such as level of education (development and database) and cultural context. There is wide variation within the human population and it is equally so among processors evidenced by the considerable qualitative difference in performance between GPT-2 and GPT-3. The experiments provided considerable evidence in support of the hypothesis that human intelligence is modelled best by learning networks exhibiting moderate degrees of lossy compression—incomplete knowledge, limited memory and error. In this sense the attributes of the artificial network provided insight into the functioning of the human network. Similarly, analysis of the responses of GPT-2 to text prompts and several of DALL-E2s artistic products provided strong evidence towards the conclusion that artificial intelligence was indeed capable of understanding natural language. Thus the second hypothesis, that the definitions of concepts such as intelligence and understanding are being broadened by such research is supported.

At this time considering the results of transformers such as GPT-3 and DALL-E 2 in tandem may be more useful than the pursuit of artificial general intelligence. The considerable artificial intelligence that has so far emerged is the result of training on vastly more data than any single human could experience in a lifetime but over a much shorter time period and without the memory of responses generated or cultural reinforcement. Future research should investigate what happens when the results of experiments such as *Madeleine* are re-entered into the learning data set of the transformer itself. In other words, it would

be fruitful to investigate the response patterns that might emerge when the transformer remembers its own history (which is the output of a second-degree social interaction with a human) and learns from it. In combination with reward and reinforcement, this might be the nexus of an even stronger convergence between machine and human cognition or perhaps further the emergence of a different kind of ‘self-awareness’ in the machine. We are no longer within the realm of science fiction; it is no more a question of either cooperation or antagonism between man and machine as separate entities. Each is already inextricably and deeply a part of the other, engaging not in a common language but beginning to create, in the words of Donna Haraway, “a powerful infidel heteroglossia”.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The GPT-2 application used was ‘talk-to-transformer’. The DALL-E 2 application used was ‘pixray/text2image’ using vqgan render engine, available on the Replicate platform. ‘Madeleine’, the robot, is produced by Abyss Creations in San Marcos, California.

Conflicts of Interest: The author declares no conflict of interest.

Note

- ¹ For a deeper discussion see: Dale Markowitz. May, 2021. Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5. <https://daleonai.com/transformers-explained> (accessed on 1 July 2021).

References

- Abraham, Diya, Ben Greiner, and Marianne Stephanides. 2021. *On the Internet You Can Be Anyone: An Experiment on Strategic Avatar Choice in Online Marketplaces*. MUNI ECON Working Paper 2021-02. Brno: Masaryk Uni.
- Aleksander, Igor. 2022. From Turing to Conscious Machines. *Philosophies* 7: 57. [CrossRef]
- Atanasoski, Neda, and Kalindi Vora. 2019. *Surrogate Humanity: Race, Robots, and the Politics of Technological Futures*. Durham: Duke University Press.
- Baudrillard, Jean. 1990. *The Transparency of Evil: Essays on Extreme Phenomena*. Translated by James Benedict. London and New York: Verso, p. 51.
- Bergen, Hilary. 2016. ‘I’d Blush if I Could’: Digital Assistants, Disembodied Cyborgs and the Problem of Gender. *Word and Text* VI: 95–113.
- Boden, Margaret. 2003. *The Creative Mind: Myths and Mechanisms*, 2nd ed. London: Taylor and Francis.
- Boden, Margaret. 2016. *AI: Its Nature and Future*. Oxford: Oxford University Press.
- Caucheteux, Charlotte, and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology* 5: 134. [CrossRef]
- Chella, Antonio, and Riccardo Manzotti. 2009. Machine Consciousness: A Manifesto for Robotics. *Journal of Machine Consciousness* 1: 33–51. [CrossRef]
- Cheng, Mingyong. 2022. The Creativity of Artificial Intelligence in Art. *Proceedings* 81: 110.
- Costa, Pedro. 2018. Conversing with Personal Digital Assistants: On Gender and Artificial Intelligence. *Journal of Science and Technology of the Arts* 10: 2. [CrossRef]
- Damasio, Antonio. 2021. *Feeling and Knowing: Making Minds Conscious*. New York: Penguin.
- DeVoss, Danielle. 2000. Rereading Cyborg (?) Women: The Visual Rhetoric of Images of Cyborg (and Cyber) Bodies on the World Wide Web. *Cyberpsychology and Behavior* 3: 5. [CrossRef]
- Fein, Daniel. 2022. DALL-E 2.0, Explained. Towards Data Science. Available online: <https://towardsdatascience.com/dall-e-2-0-explained-7b928f3adce7> (accessed on 1 June 2022).
- Goh, Gabriel, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. *Distill* 6: e30. [CrossRef]
- Haraway, Donna. 1991. *Simians, Cyborgs, and Women: The Reinvention of Nature*. London: Routledge.
- Korteling, J. E. Hans. 2021. Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence* 4: 622364. [CrossRef]
- Lamba, Harshall. 2019. Intuitive Understanding of Attention Mechanism in Deep Learning. Available online: <https://towardsdatascience.com/intuitive-understanding-of-attention-mechanism-in-deep-learning-6c9482aecf4f> (accessed on 1 June 2022).
- Machneva, Maria, Anthony M. Evans, and Olga Stavrova. 2022. Consensus and (lack of) accuracy in perceptions of avatar trustworthiness. *Computers in Human Behavior* 126: 107017. [CrossRef]

- Maguire, Phil, Philippe Moser, and Rebecca Maguire. 2016. Understanding Consciousness as Data Compression. *Journal of Cognitive Science* 17: 63–94. [[CrossRef](#)]
- Manning, Christopher D., Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* 117: 30046–54. [[CrossRef](#)]
- Masterman, Margaret. 2003. *Language, Cohesion and Form: Selected Papers of Margaret Masterman, with Commentaries*. Edited by Yorick Wilks. Cambridge: Cambridge University Press.
- Nyiri, J. 1989. Wittgenstein and the Problem of Machine Consciousness. *Grazer Philosophische Studien* 33: 375–94. [[CrossRef](#)]
- Omry, Keren. 2005. A Cyborg Performance: Gender and Genre in Octavia Butler. *Praxis Journal of Philosophy* 17: 45–60.
- Pepperell, Robert. 2022. Does Machine Understanding Require Consciousness? *Frontiers in Systems Neuroscience* 16: 52. [[CrossRef](#)]
- Piletsky, Eugene. 2019. Consciousness and Unconsciousness of Artificial Intelligence. *Future Human Image* 11: 66–71. [[CrossRef](#)]
- Quiroga, R. Quian, Leila Reddy, Christof Koch, and Itzhak Fried. 2007. Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of Neurophysiology* 98: 1997–2007. [[CrossRef](#)]
- Schmidhuber, Jürgen. 2009. Simple Algorithmic Theory of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes. *Journal of the Society of Instrument and Control Engineers* 48: 21–32.
- Schmidhuber, Jürgen. 2010. Formal Theory of Creativity and Fun and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2: 230–47. [[CrossRef](#)]
- Sutherland, Stephani. 2015. What Happens in the Brain when we Read? *Scientific American: Mind* 26: 14. [[CrossRef](#)]
- Vallor, Shannon. 2021. The Thoughts the Civilized Keep. *Noema Magazine*, February 2.
- Vogl, Elisabeth, Reinhard Pekrun, Kou Murayama, Kristina Loderer, and Sandra Schubert. 2019. Surprise, Curiosity, and Confusion Promote Knowledge Exploration: Evidence for Robust Effects of Epistemic Emotions. *Frontiers in Psychology* 10: 2474. [[CrossRef](#)] [[PubMed](#)]
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations I*, 281,395f. Oxford: Blackwell.