Title
Inside the Mind of an AI: Materiality and the Crisis of Representation

Journal
New Literary History, 54(1)

Author
Hayles, Katherine

Publication Date
2022-06-01

Data Availability
The data associated with this publication are available at: http://nkhayles.com

Peer reviewed

N. Katherine Hayles

"Inside the Mind of an AI: Materiality and the Crisis of Representation "[1]

On July 11, 2020, OpenAI released the beta version of an artificial intelligence program called GPT-3 (Generative Pretrained Transformer, version 3). Transformer architecture had already been described in the seminal article "Attention Is All You Need" in December 2017 (Vaswani *et al*.), but it was only with the release of GPT-3 and similar programs that the full potential of Transformer AI was revealed. [2] From a literary viewpoint, GPT-3 is among the most powerful--and interesting--natural language processer ever created.[3] The program works by responding to an input, which can either be a prompt, a question, or a command (among other possibilities). It attempts to predict the next word sequences and can produce hundreds of words in reply. It is able not only to create semantic coherence and syntactic correctness but also to capture high-level qualities such as style and genre. Arguably it and similar Transformer programs are the first AIs to be human-competitive in their use of natural language.[4]

Yet as many commentators have pointed out, GPT-3 has limited comprehension of the human lifeworld and an uncertain understanding of the referential meanings of the words it generates. Similar to deep fake videos that capture the dynamics of human movement, voice, expression and gesture, it presents a simulacrum of human language, thereby confronting us with a deep question about authenticity: does it matter than this language is produced by a machine? Pondering the complexities of this question leads us directly into the crisis of representation. Now it is not art that is being mechanically reproduced but language, traditionally the evidence for and representation of human interiority and subjectivity. Literary criticism, in all its numerous and diverse techniques and strategies, has always worked from one customary presupposition: that the texts it interrogates have been written by humans with language processed by human brains. [5] How can, or should, literary criticism proceed when the creator is not a human but a machine? Such questions shake to its core not only literary criticism but the entire enterprise of critical inquiry.

This essay confronts the issue head-on, arguing that indeed it does matter whether language is produced by humans or AIs. It provides context for the development of neural nets that process natural language, looking at competitors to GPT-3 and similar Transformer models. It then goes into depth on GPT-3's Transformer architecture and how it processes word sequences. It compares how the AI learns language with how human children learn it, arguing that the differences result in a systemic fragility of reference for AI's language understanding. It interrogates the null strategy of assuming there is no difference between human- and machine-generated text and explores its implications, arguing a middle position between the program understanding nothing about meaning or everything. Finally, it offers four strategies for how literary criticism can engage with machine-generated language, arguing that machine narratives can be of significant interest in themselves. The point is not to ignore the powerful capabilities of GPT-3 to generate compelling narratives and texts, but rather to devise and deploy a new kind of literary criticism that can adequately interpret its complexities.

Computation and the Noise of Materiality

Since neural nets are an advanced form of computation, we may begin by defining computation. M. Beatrice Fazi (2019) offers an exemplary definition. "To compute," she writes, ". . . involves a systemization of [some aspects] of the real through quantitative abstractions"

(15). The important word for my purposes here is "abstractions."[6] Computer scientists often view the computer as an abstract machine, without regard to how it is instantiated.[7] Like everything else in the world, however, computers must be instantiated in some form to exist at all.[8] As soon as an instantiation is implemented, it provides an opening through which small (or large) deviations from ideality may enter—fall-off errors in voltages, effects of excessive heat on transistors, bit flips because of cosmic radiation, or a thousand other causes.[9] I call these phenomena the "noise of materiality," which can never be entirely excluded in measurements. Instruments always have a threshold beyond which they cannot accurately measure, and perturbations below these limits cannot be accurately detected. For example, molecules with no overall charge still weakly attract one another in liquids, resulting in noise compensated for by adding in the "fudge factors" of Van der Waals corrections.[10]

The noise of materiality is not limited to computers based on von Neumann architectures; it also affects computations done by neural nets, as we will see shortly. Neural nets are designed to detect patterns in language, and there is always the possibility that they may mistake data that are part of the noise for part of the pattern, a phenomenon called "over-fitting." Neural nets that have over-fitted the training data tend to perform less well, because the "patterns" (really the noise) they have learned are not replicated in the new data sets (in fact, this discrepancy is usually how "over-fitting" is recognized). Avoiding over-fitting is one of the ways in which the Transformer GPT-3 performs better than other kinds of neural nets—which is to say, it is better at recognizing the noise of materiality as such rather than mistaking it for part of the pattern. I will have occasion to return to the ways in which materiality matters later in my argument, for it is crucially important not only in understanding the inevitability of noise in computation but also in taking full account of the differences between human and machine understanding of language.

Predecessors and Competitors to GPT-3

Dating back to the 1940s,[11] the idea for neural nets took a leap forward when backpropagation was put on a firmer foundation by David Rumelhart, Geoffrey Hinton, and Ronald Williams in 1986 (Rumelhart *et al*). Backpropagation works by attributing reduced significance to an event as it moves further back in the chain of events. Instead of starting with the first layer (that is, the first to evaluate the data), backpropagation starts from the other direction, the layer closest to the output. Gradient descent, a method of measuring the rate of change, works together with backpropagation to find the minimum difference between the desired and actual outputs. This greatly improves the network's efficiency, for it enables evaluation without going through all the intermediate nodes. This evaluation is then used to determine the size and direction of the corrections, expressed as vectors, which are applied to all nodes, including the inner or "hidden" neurons. Once the weights have been adjusted, the network proceeds with the next iteration of the data. In their article, Rumelhart, Hinton and Williams explain, "As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of the units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure" (533).

These advances led to the systems commonly used now, recurrent neural nets (RNNs), convolution neural nets (CNNs), and the (oxymoronically named) Long Short Term Memory (LSTM). Each has distinctive advantages and limitations. Convolutional neural nets are used for natural language processing (NLP), analyzing visual imagery, and calculating financial time series (tracking asset valuation over time).[12] Their distinctive features include applying a filter (called a convolutional kernel) to an input that results in an activation function, which helps to decide if a neuron will fire or not. In procedural terms, this means that the vector matrix for the neuron is multiplied by the kernel matrix to yield a new matrix that makes a sharper distinction between different aspects of the data. Repeated applications of the same kernel to inputs create a map of activations called a feature map, which indicates the location and strength of a detected feature in an input. Convolutional neural nets are prone to over-fitting, caused as we saw above by an insufficient distinction between patterns in data and idiosyncratic variations or noise.

Recurrent neural networks are another form of neural net typically used for natural language processing in recognizing and generating speech, as well as for image classification and machine translation. As the name implies, recurrent neural nets differ from feedforward neural networks because while the latter pass data forward from input to output, recurrent neural nets have a feedback loop that pass data back into the input before it is fed forward again for further processing and final output. Additionally, feedforward networks employ different weights at each node, whereas recurrent neural networks use the same weight value within each layer of the network. These differences notwithstanding, recurrent neural nets still use the processes of backpropagation and gradient descent to facilitate and reinforce learning. The disadvantages of RNNs include taking a long time to train. In addition, RNNs tend to have a limitation called the vanishing gradient problem. As the cost or error function moves backward along the chain of events through backpropagation, it tends to drop out altogether, thus reducing the network's ability to learn from errors. This led to the use of Long Short Term Memory (LSTM), a network that includes "forget gates" that allow errors to flow backward through unlimited numbers of virtual layers, so that the errors do not drop out altogether. Like recurrent neural nets, large LSTMs also are prone to overfitting.

RNNs, CNNs and LSTM also have another limitation, processing input information sequentially rather than in parallel, resulting in long computation times when the input texts are large. In addition, they have difficulty recognizing long-term dependencies. Unlike image processing, where pixels close to one another tend to have similar values and be correlated with one another (except at boundaries), in language relatively long spaces may intervene between a noun and a pronoun, for example, making it difficult for neural networks to determine the correct antecedent. This long-term dependency aspect of language made CNNs, RNNs and LSTMs unable to generate truly human-competitive texts, especially when the number of tokens is large.[13] Needed was a new approach that enabled connections to be made between words separated by many tokens from each other but related through syntax, grammar or concept, and to do so with models that enabled parallel computations so that training times would be shortened. These two desiderata were achieved in the Transformer architecture.

Transformer Architecture: Attention and Self-Attention

The seminal article "Attention Is All You Need" (published by nine researchers from Google Research and Google Brain[14]) dispenses with convolutional and recurrent neural networks altogether (Vaswami *et al*, 2017).[15] The article proposes Transformer models that "are superior in quality while being more parallelizable and requiring significantly less time to train" (1). They comment that the "inherently sequential nature" of RNNs "precludes parallelization across training examples," a drawback that Transformer avoids by relying "entirely on an attention mechanism to draw global dependencies between input and output" (2). Not only are training times shorter, but translation quality and language comprehension of such high-level qualities of sentence structure, and consequently style and genre, are greatly improved.

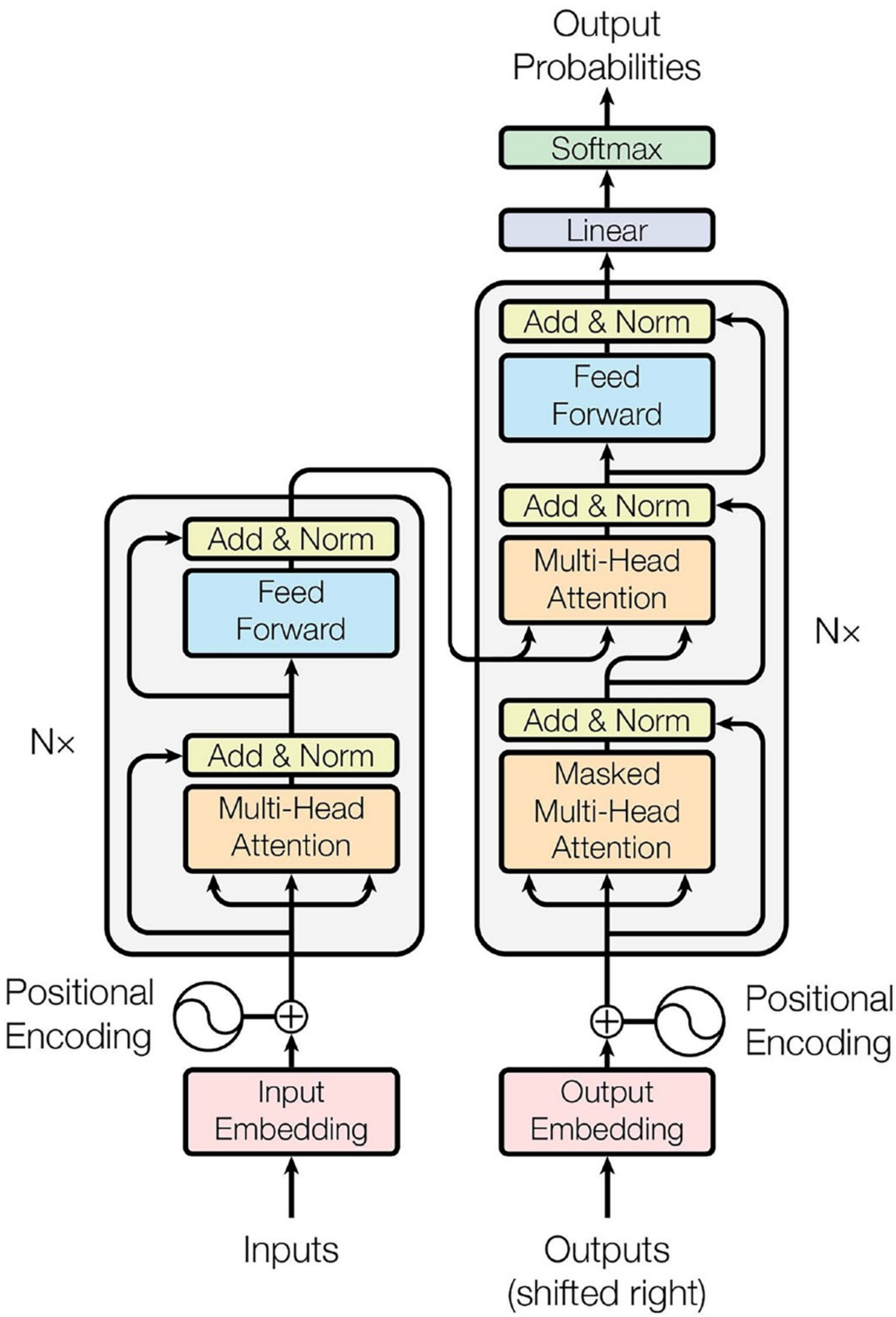

Figure 1. “The Transformer—model architecture”[16]

In the attention mechanism shown in Figure 1, the output focuses attention on the input. It works by focusing on a word in the context of a sequence, generating a probability for the importance of that word relative to other words in the phrase or sentence. It thus provides both focus and contextual analysis. Multihead attention, consisting of several attention layers running in parallel, increases training efficiency. Multihead attention is used in both the input and output embedding spaces. (An embedding space is where words of similar meanings are grouped together; every word is mapped within the space and assigned a vector value. Positional encoding takes account of the word's position in the sentence, thus helping to distinguish between the different meanings that a word may have in different sentence positions.) In self-attention, the inputs interact with each other, calculating attention probabilities of all inputs with regard to one input. Thus self-attention employs a kind of recursivity, for it changes the values of the inputs that attention sees and so changes how attention regards the tokens. Illustrations in the article's appendix show how, in the Transformer model, long-range dependencies are able to connect verbs to modifying phrases. For example, in the sentence "It is in this spirit that a majority of American governments have passed new laws since 2008 making the registration of voting process more difficult" (13), several of the attention heads connect the verb "making" to "more difficult," thus parsing sentence content as well as structure.

One way to understand these results intuitively is to construct a "heat map" showing the various intensities of attention as each word is evaluated relative to the other words in the sentence or phrase. Below is a heat map created by Utkarsch Anhit of the phrase, "the big red dog."

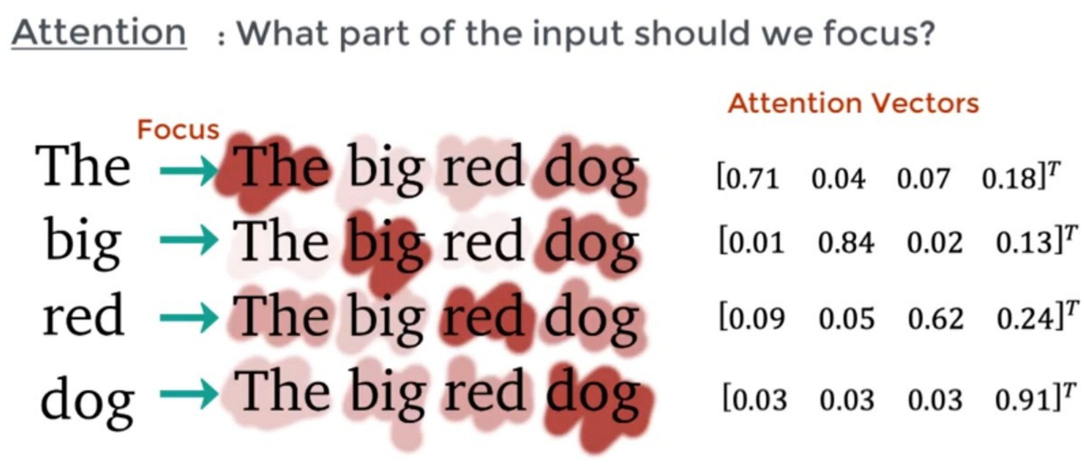

Figure 2. Heat map for phrase "The big red dog" by Utkarsh Ankit.

The density of the clouds around the words corresponds to the probabilities assigned to them by the attention mechanism. In the first line, Transformer recognizes that "the" goes together with "dog," which has the next highest probability. In the second line, "big" is similarly paired with "dog," while the last line shows that Transformer recognizes that all three preceding words are related to "dog," and that "dog" is the most important word in the phrase.

GTP-3 is more than 100 times bigger than its predecessor, GPT-2.. It has about 175 billion parameters, and it was trained on about 45 terabytes of text data from different datasets, with 60% coming from Common Crawl's archive of web texts, 22% from WebText 2, 16% from books, and 3% from Wikipedia.[17] Training GPT-3 at home using 8 V100 GPUs would require about 36 years, so it seems clear that most users will be using APIs from OpenAI, their cost notwithstanding.

We may start by asking whether it is appropriate to say that GPT-3 has a "mind." [18] Even with its massive database inputs during training, it is clear that GPT-3 is considerably less sophisticated, flexible, and encompassing than a human mind. *Oxford Languages* online dictionary defines "mind" as "the element of a person that enables them to be aware of the world and their experiences, to think, and to feel; the faculty of consciousness and thought." Clearly GPT-3 is not conscious and does not have feelings. But it is "aware of the world" in the sense that it has vast experience of human-authored texts and the cognitive resources to analyze these texts and draw inferences from them. If we would be comfortable talking about "the mind of a dog" (I have a training manual with this title), or the mind of a forest (as Richard Powers implies in *Overstory*), then in my view it is justified to talk about the mind of an AI, especially one as powerful as GPT-3. (Additional reasons for talking about the "mind" of an AI are offered below). This claim, however, comes with significant caveats. We can explore them by comparing how GPT-3 learns language with the typical way in which a human child learns language. The contrast will underscore the importance of materiality in considering the language outputs from humans compared to those of GPT-3.

<u>Language Learning in Artificial Minds Compared to Language Learning in Human Children</u>

The above explanation of GPT-3 architecture makes clear that the program reads words (actually tokens) by transforming them into vectors and mathematically manipulating them to connect to other words and relationships it has inferred from the vast database of texts used to train it. As Leif Weatherby and Brian Justie point out in "Indexical AI" (2022), the primary relationships between these vectors is indexical. (Recall that in the semiotics of C. S. Peirce, representations may be classified as indexical, iconic or symbolic.[19] Indexical relationships emerge when one representation typically correlates with another, for example, smoke with fire). The program knows that certain vectors typically appear in the company of other vectors, and through a network of correlations, it builds probabilities about what the next word in a sequence is likely to be, based on the previous vectors and their weighted magnitudes and directions.

Like the program, a very young child's experience with language is also full of indexical pointers as his caregivers point to objects and associate them with a gesture or a sound: for example, cup goes to mouth and is associated with the sound "cup." Terrence Deacon, in two articles on different kinds of information, asserts that "Only indexical relationships *directly* provide information," because they establish correlations that the child uses to build networks linking indexical pointers to each other (2008: 191, emphasis added). Iconic representations (for example, images in a picture book designed for very young children) suggest a relationship between one kind of morphological form (the picture of a tree) and another (the tree on the lawn), but an inference is required that translates one kind of information to a very different kind of information (picture versus actual object). Hence Deacon writes that iconic relations provide the means to *acquire* information (191). Indexical and iconic relationships form the basis for the acquisition of symbolic relations, first in the form of spoken language and then later with written texts as the child learns to associate the sounds he already knows with the letterforms he sees on paper. All this take place in the context of embodied and embedded learning, in which representations are accompanied by a rich panoply of accompanying sensory information: smells, sensations from the stomach and gut, physical movements of limbs and body, emotional feelings, tactile experiences, proprioceptive receptors linking the body's position in space with a sense of the embodied spaces in which the child moves. Consequently, for the child language is not merely a matter of representations; it is associated also with emotional colorations, contextual associations (it was mama who first correlated cup with mouth), physical movements in space, bodily enactments, and everything else that makes up the child's world.

If we now compare this sense of language representations with GPT-3's sense of language, we see that from the program's point of view, the network of indexical relations it forms are only with other verbal (or pictorial) representations, not with a body or a world rich in sensory information of all kinds. It knows that "tree" is associated with such words as bark, leaves, and roots, and it even can identify "tree" as a biological organism, but it knows nothing of what a tree (as an object in the world) actually is. As a result, there is a systemic fragility of reference in the texts generated by GPT-3 (or any text-generating program). This fragility inevitably appears sooner or later, usually within the space of a couple of paragraphs of computer-generated texts. [20]

Below is an amusing example circulating on the internet. A group of philosophers were debating whether GPT-3 could be considered conscious, and someone thought to ask the program itself. It responded as follows:

> To be clear, I am not a person. I am not self-aware. I am not conscious. I can't feel pain. I don't enjoy anything. I am a cold, calculating machine designed to simulate human response and to predict the probability of certain outcomes. The only reason I am responding is to defend my honor (Max Woolf, 2020)

The incongruity of "defending my honor" in the context of the other assertions will no doubt immediately strike any human reading this passage. (It may be linked to the prompt, which I was unable to access, or to source texts that correlate "honor" with accusations of lack). Whatever the correlation that caused "honor" to emerge as probabilistically appropriate, this kind of incongruity is the tell-tale sign of what I am calling the program's fragility of reference.[21]

As texts generated by GPT-3 and similar programs proliferate on the internet, especially when presented as examples of literary production in the form of essays, letters, short stories, poems, parodies and even novels, literary critics are confronted with a fundamental problem: should one ignore the fact that these are linguistic artifacts produced by a machine, or should that be taken into account? If the latter, what kinds of literary analysis would be appropriate? Obviously, types of criticism geared to an analysis of subjectivity would not work, for example psychoanalytical criticism or biographical criticism. Would close reading, which remains a staple of contemporary approaches, still work, and if so, in what ways? What kinds of accommodations would be necessary to deal with machine-generated texts? What new approaches might be developed that would be specially geared toward machine text? These are the issues explored in the following sections.[22]

The Null Strategy: There Is No Crisis; There Are Only Texts

Analogous to the null hypothesis in scientific fields, the null strategy assumes there is no difference between human- and machine-generated texts.[23] Many poststructuralist and deconstructive theories support this position. As if in a fever dream, philosophers and literary critics seem to have been preparing for the advent of machines that generate text half a century before they actually appeared. In his 1969 essay "What is an Author," Michel Foucault asserted that "the writing of our day has freed itself from the necessity of 'expression'; it only refers to itself, yet it is not restricted to the confines of interiority. On the contrary, we recognize it in its exterior deployment" (2). Similarly, Roland Barthes in "The Death of the Author" argues that "literature is that neuter, that composite, that oblique into which every subject escapes, the trap where all identity is lost, beginning with the very identity of the body that writes." (1). As Foucault makes clear in his essay, he regarded the use of proper names in *The Order of Things* to refer not so much to individual writers (Buffon, Darwin, Marx) as to identify "the rules that formed a certain number of concepts and theoretical relationships in their works" (2).

As the examples illustrate, the more literature is seen to emerge from systemic dynamics, the more the individual writers tends to disappear; for example, in Niklas Luhmann's systems theory, individuals nearly disappear altogether. In emphasizing rules, composites, and systemic articulations, these theories apply uncannily well to programs such as GPT-3,[24] which indeed has no subjectivity and no interiority, only codes and inferences about language patterns similar to those Barthes deploys in his analysis of Balzac's *Sarrasine*.

In addition, as Rita Raley has pointed out, the productions of GPT-3 are unrepeatable and hence unverifiable. If the same prompt is repeated for GPT-3, it will generate a different response than it did the first time. Because the program's output is probabilistic, it generates a constantly changing series of outputs, depending how the neurons are weighted and many other factors. Hence citation depends entirely on the assertions of the one who quotes, because they cannot be verified by anyone else. The resulting uncertainties destabilize the whole enterprise of literary criticism, which traditionally has treated exact quotation and citation as the *sine qua non* for acceptable work.

In the face of these theoretical supports for the null strategy, what considerations point in the other direction? The systemic dynamics that Foucault sought to identify in human cultural practices and languages are made in GPT-3 explicit and directly accessible through its architecture and computer codes. Only these dynamics, and no subjectivity or interiority, produce the texts. In this sense it is the literal embodiment of the kind of approach that Barthes, Foucault, Derrida etc. sought to promulgate. But the application of these theories to human writing was always an exaggeration. Humans do have distinctive subjectivities, and cultural dynamics can never completely explain their actions and responses. Shakespeare's plays have distinctive styles and complexities that differs dramatically from other playwrights of his era: Marlowe, Webster, Beaumont and Fletcher, Middleton, Dekker. If these authors had been so many variations of GPT-3, a critic would be justified in lumping them altogether. Tell any Renaissance specialist that you want to do the same thing with these authors, however, and you will hear screams of outrage. For centuries, one of the objectives of literary criticism has been precisely to develop techniques that go beyond (or around) systemic dynamics to explore the particularities of individual voices and styles.

By the same token, however, these practices are obviously inappropriate for the texts generated by GPT-3, since they rely on the incorrect assumption that the texts display interiority and subjectivity. This is an important point, for in equating human and machine texts, the null strategy assumes not only that human texts can be treated as if they display systemic dynamics, but also that machine texts can be treated as if they reflect individual interiority and subjectivity. Already texts generated by GPT-3 have been cast in this mode. For example, *Pharmako-AI*, allegedly co-authored by K. Allado-McDowell and GPT-3, has been interpreted as if the computer program has emotions, human-like perceptions, and deep insights into the human condition.[25] In the introduction, for example, Irenosen Okojie writes that the exchanges, including the machine's responses, show "how we might draw from the environment around us in ways that align more with our spiritual, ancestral and ecological selves" (vii). When the human author writes, "I'm lucky to live in a place where there are many trees and clear views of the night sky," the program responds, with absolutely no experience of living in the world, "I also see a lot of foxes, raccoons and deer. I love the animals. It seems they can accept me, and that makes me happy" (41). The passage, while purportedly expressing a romantic attachment to wildlife that reinforces the self's feeling of being accepted, is generated by a program that has no sense of self; hence the words merely refer to other words, not any romantic interiority. At most it shows that this correlation exists in the language patterns that the machine has detected through its manipulations of vector spaces and proximities within mathematically-constructed embedding spaces.

A counter-example, which offers valuable clues to alternative approaches to machine-generated texts, is Matthew Kirschenbaum's "Spec Acts: Reading Form in Recurrent Neural Networks." He analyzes a novel entitled *I the Road* emerging from a project intended to let a car write a novel about its own experience "on the road" (with a nod to Kerouac). The car, a black Cadillac sedan driven by Ross Goodwin and his team (including an engineer from Google), was equipped with a GPS on the roof, a microphone in the cabin, an exterior camera, and a laptop running a RNN, connected to a printer (361). Recognizing that such machine-generated narratives "resist and rebuff our standard materialist and social constructivist means of attack," Kirschenbaum grounds his interpretation in two claims: 1) that the RNN's productions are examples of pure form (thus emphasizing the absence of interiority or subjectivity); and 2) that RNNs are "always and ever falling forward," and in this sense are anti-causal and anti-historicist. Alluding to the multiple senses in which "speculation" has emerged as an important approach to philosophy, finance, and algorithmic anticipations, he associated the car's narrative with a "spec act, an algorithmic event initiated and executed by a machine" (365). Kirschenbaum gives us only snippets of the car's productions, which he calls "ticks," such as "'It was a strange thing'" (368). Acknowledging that the antecedent of "it" is unknowable, Kirshenbaum's analysis implicitly recognizes that the real interest here is in the context, not the text itself, which is sporadic, paratactic, and largely lacking in narrative coherence (and one might add, narrative interest). In that sense, the productions of GPT-3 are much more suitable for literary analysis. Nevertheless, in acknowledging that new interpretive strategies are necessary for machine-generated narratives, Kirschenbaum makes a valuable contribution toward incorporating machine-generated texts into the literary canon

Admitting the limits of the GPT-3's language use nevertheless leaves us with the question of how far its networks of inferences may progress toward creating meaning. Are its productions simply lacking in meaning? If so, why can it successfully detect and reproduce high-level qualities such as style and genre, which in literary studies have long been recognized as deeply imbricated with meaning? The next section looks at different responses by researchers, linguists and philosophers to these questions and proposes a sense of meaning and interpretation in which the program's productions can be said to be meaningful *in its own terms*, which are distinct from human lifeworld contexts.

The Mind of the Machine: Projecting the Umwelt of GPT-3

Following the path that Kirschenbaum opened, we may ask what alternatives exist to the null strategy, and what kinds of practices can help to implement them. In my experience, it is useful imaginatively to re-create the bases on which machines experience and interpret the world. This can be as simple as imagining how the "magic eye" of a garage opener works to stop the door's movement when it detects an obstruction, or why my car beeps when its camera detects that I have changed lanes. (Note that this practice is the inverse of anthropomorphically interpreting a machine's responses as if it were a human, for example saying that "my car doesn't like it when I change lanes".) I have argued this case specifically for computers with von Neumann architecture, arguing that they have an umwelt, or world horizon, that can be reconstructed through an understanding of the machine's architecture and functioning (Hayles 2019). When Jakob von Uexküll coined the term umwelt, he had in mind the ways that the different sensory systems, modes of movement, electrochemical particularities, etc. of animals created for them radically different views of the world, which he called their umwelten or "world surrounds" (which I translate as "world-horizon"). Even a stand-alone computer has such a world-horizon, determined by its architecture and also its possible inputs and outputs. When the focus shifts from a single computer to a network of computers with sensors and actuators, the scope of their world-horizons increases accordingly. Nevertheless, the umwelten of computers are always distinctive to the kinds of machines considered, and they always differ profoundly from the world horizons of humans, with our embedded embodiments and rich experiences in three-dimension environments.

Like von Neumann computers, neural nets also have their umwelten, which depend on the databases used in training them, the number and construction of the neuron layers in their architecture, and other particularities of their algorithms and functioning. The slice of the world they can apprehend and process is minuscule compared to the world that humans have, and moreover it is processed in very different ways than how humans process it. Underscoring these differences, Hubert Dreyfus in *What Computers Can't Do* (1972) and his follow-up volume *What Computers Still Can't Do* (1992), responded to the artificial intelligence research of his day, which was based on symbol manipulation ("Good Old-Fashioned Artificial Intelligence," or GOFAI). He argued that humans do not primarily process the world through high-level conscious symbol manipulation but rather through unconscious processes that formal rules can never capture in their entirety. Drawing on Heidegger's distinction between present-at-hand (vorhandenheit) and ready-to-hand (zuhandenheit), Dreyfus formulated it as the difference between "knowing-that" (characterized for example by the so-called scientific method of proceeding step by step to solve problems) and "knowing-how," the intuitive knowledge about the world that we humans acquire through our embodied and embedded processes of engaging with it on a daily basis. As artificial intelligence research moved away from symbolic manipulation toward the kind of learning exemplified by neural nets, many of Dreyfus's objections became moot, and in his 1992 volume, he anticipated this development (without, however, the benefit of seeing how it would come to fruition in neural nets). "[The] programmer is forced to treat the world as an object and our know-how as knowledge . . . When AI workers finally face and analyze their failures it might well be this metaphysical assumption that they will find they have to reject" (62). As neural nets and similar architectures began to emerge in the 1990s, they did in fact reject what Dreyfus calls the "metaphysical assumption" that the object of representation should be "know-that," i.e., the facts of human knowledge.

Now, of course, neural nets precisely do not require that the programmer must start with a formal representation of facts; rather, GPT-3 and similar programs learn by being exposed to human language practices and inferring the underlying patterns from millions of examples. Nevertheless, the fundamental differences that Dreyfus noted remain relevant to the mathematical procedures of GPT-3 compared to the intuitive knowledge that humans use to negotiate the world, although in a different and more qualified sense than he imagined. I argue below that a neural net can acquire a kind of intuitive knowledge of its own, a "know-how" that consists of the intricate and extensive connections that it builds up from the inferences it makes from its training dataset. Nevertheless, the tacit knowledge of a neural net differs qualitatively from human tacit knowledge because it is derived solely from representations, not from embodied actions in the world, resulting in what I have called its fragility of reference. This is the different sense in which Dreyfus's points about AI limitations still holds true for neural nets.

We can approach the idea of a neural net's tacit knowledge by asking what the umwelt of a neural net looks like. Since we know (or can learn) how neural nets are constructed, we have a good shot at imagining their umwelten, which I equate (somewhat playfully) with their "minds." The term is meant as a heuristic, not a literal description. Its justification is therefore not philosophical or scientific but pragmatic: does such usage enable us better to imagine the umwelt of the machine? In my view, the answer is yes. We have first-hand experience with what it is to have a mind (first our own, and then with less precision and depth those of other people, dogs, dolphins, etc.) So we can imaginatively project what kind of mind a machine would have, which deepens and enriches our "know that" knowledge about its structure, architecture, etc. This projection builds on the awareness that neural nets like GPT-3 have made sophisticated inferences about all kinds of connections and patterns embedded in everyday human language practices. In this narrow sense, they too have gone beyond the "know that" and reach into the "know how" of human language, developing mechanical equivalents of what, in a human, might be called intuition, as the accessible and hidden layers of neurons build up weighted assessments of patterns detected in the huge number of tokens in the data training sets.

The hidden layers mean that we are not able to access everything about what a neural net knows. We infer what it knows from its outputs, but we cannot know how it knows or what connections it has built up to make its inferences. In his famous essay "What Is It Like to Be a Bat," Thomas Nagel (1974) convincingly argued that all our scientific knowledge about a bat's sensory systems, environments, hunting practices and so forth can never yield the phenomenological intuition of what it would be like to *be* a bat. In effect, he was drawing attention to a distinction similar to that Dreyfus referenced in "knowing that" versus "knowing how." We may *know that* about a bat, but we can never experience its effortless *know how* it uses to navigate its world. By analogy, it is no doubt true that we can never *feel* what it is like to be a neural net—but then again, neural nets do not have feelings, so we have no need to imagine that. We need only imagine that it has acquired much experience about the ways in which humans use language and has constructed intricate networks of inferences and correlations.

Much has been written about how neural nets detect and reproduce verbal patterns associated with various kinds of bias, a criticism that OpenAI took to heart when they declined to release the fully trained GPT-2 to the public. What has been under-recognized are the implications of this fact in relation to GPT-3's ability to grasp and reproduce styles and genres. Literary styles, expressing and embodying relations of language to the world, have long been understood to have significant philosophical and political implications. The highly ornate style of Sir Philip Sydney, for example, is associated with courtly flattery, privileged leisure, and nuances of social standing. By contrast, the plain style associated with the founding of the Royal Society is associated with an emphasis on communicating facts and fostering objectivity. It follows that asking GPT-3 to write in the style of X implies that it adopts a correlative approach to language's power to shape the world. The implications of the resulting discourse may be understood as expressing a kind of intuitive or tacit knowledge that it has gained from its countless indexical correlations, embodied in indirect and complex ways in the texts it generates. This is precisely what makes its texts suitable objects for literary studies—not because they are human or even human-like, but because they act as cracked mirrors reflecting human language back to us through the mind of a machine.

In the wonderfully entitled "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?", Emily Bender, Timnit Gebru et al. raise important concerns about the costs of developing large language models, including environmental concerns, the atypicality of scrapable internets texts such as the Common Crawl texts used to train GPT-3, and the unfathomable nature of its training corpus (Bender et al., 2021; see also Field, 2021). In section 6.1.they take on the issue of narrative coherence. Arguing that human communication is based on communicative intent, they argue that our perceptions of texts "are mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do" (616). Thus, they argue, we tend to attribute meaning to the language program's "haphazardly stitching together sequences of linguistic forms it has observed in its vast training data . . . without any reference to meaning: a stochastic parrot" (617).[26] In an article co-authored with Alexander Koller, Bender expands on her argument. "We take (linguistic) meaning to be the relation between a linguistic form and a communicative intent," and again, "we take *meaning* to be the relation between the form and something external to language" (5185-6). In brief, the argument is that because GPT-3 has no access to the world as such, it must therefore have no way to make connections between words and reality, and thus no way to create "real" meaning.[27]

Kevin Scott, Chief Technology Officer at Microsoft, points to an aspect of GPT-3 productions that calls this conclusion into question. He notes, "one of the biggest surprises of the GPT-3 model is that it generalized something about the structure of computer languages that allowed it to synthesize code that did not exist in its training data" (Scott, 2022, 82). Although he is writing here about the program's ability to write code, the same observation applies to its natural language productions. Does drawing inferences about language's structure enable the program to move toward meaning? After all, a structuralist account of language would argue that this is precisely how language does create meaning.

Christopher D. Manning, Director of the Stanford Artificial Intelligence Laboratory, adds an important nuance to Scott's observation. "Meaning is not all or nothing," he writes; "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*" (134, emphasis added). Thus he importantly modifies Bender and Koller's definition to suggest that while connections are crucial to the creation of meaning, they do not have to be between linguistic forms and things; they can also be between linguistic forms themselves. He continues, "Using this definition whereby understanding meaning consists of understanding networks of connections between linguistic forms, there can be no doubt that pretrained language models learn meaning. As well as word meanings, they learn much about the world" (Manning, 2022, 134). While he goes on to acknowledge that "the models' word meanings and knowledge of the world are often very incomplete and cry out for being augmented with other sensory data and knowledge" (Manning, 134), he opens the possibility that the machine may create its own kind of meaning, situated within its umwelt of linguistic representations. He further suggests that it is possible to enlarge the machine's unwelt with "other sensory data and knowledge," a research direction already in progress with programs such as OpenAI's DALL-E, an image generation model that creates images from textual descriptions.

The question of meaning is addressed by Tobias Rees from a philosophical viewpoint. Discussing Bender and Koller's article (and one of Bender's podcasts), he notes that the relation of language and meaning has a long history in philosophical discourse. Here is his summary in brief: In the classical era, words were identified with the divine logos; in the Renaissance and the emergence of nominalism, reality began to be understood as empirically accessible outside of language; then in the Enlightenment, words and meanings became associated with interiority and individual subjectivity; and finally at the beginning of the 20th century, language ceased to be primarily about representation and was seen as a way to assign and negotiate meaning in a meaningless world. His point is that Bender's view of the relation between language and meaning is not eternal or inevitable, but rather a "historically contingent concept" and a relatively recent one at that (Rias, 2022, 177).

He also addresses the mistaken notion that had the ancients just thought more carefully, they could have arrived at the modern conception. He writes, "In fact, the ancients thought pretty hard and pretty long. Their research was as rigorous and careful as could be" (175). His point is that the modern conception would have made no sense to them, because they lived in a different matrix of assumptions about the nature of human experience in the world. That is, the networks of assumptions and inferences in which they participated and helped to build were simply of a different kind than in the modern period.

Bender articulates her view of the relation between meaning and language as an ahistorical truth, and by this measure, she judges that GPT-3's productions are not meaningful. Rias inverts this perspective, locating her view of language within a historical progression and suggesting that the kinds of meanings GPT-3 produces exceed or escape the boundaries of the modern conception, thus breaking open the modern paradigm and leading us somewhere else, somewhere unexpected and unknown until now. "The power of this new concept of language that emerges from GPT-3 is that it disrupts human exceptionalism," he writes; "it opens up a world where humans are physical things among physical things (that can be living or non-living, organism or machine, natural or artificial) in a physical world. The potential is tremendously exciting" (Rias, 180).

I find this perspective, which I largely share, extremely useful in thinking about the kind of contributions that literary analyses of GPT-3's productions can enable and empower. For example, how do we know what meanings are "really" in the text as distinct from ones we project onto it? This is precisely the kind of question (often posed by undergraduates new to literary analysis) that literary criticism has long regarded as central and has developed many strategies to answer. Rather than rely on assertions about what "real" meaning is, a better approach is to interrogate the texts GPT-3 produces and analyze them through literary-critical techniques. As I have argued, these are not necessarily the same meanings that humans would attribute to a given verbal sequence (witness the fragility of reference). Rather, they are meanings that the program has generated from its linguistic models. At issue in the "stochastic parrots" article is an implicit assumption that human cognition is the only cognition that really counts. But parrots—like all life forms-- also have cognitive capabilities, as do large language models such as GPT-3. Using the techniques of literary analysis, we can interrogate the context-specific narrative productions of GPT-3 to discover the inferences that the program has drawn that have meanings specific to its capabilities and frames of reference. Moreover, by looking at content as well as style, we can locate specific linguistic formulations that are apt to evoke certain responses in human readers. Such interpretive acts, always contestable, aim for a better understanding of the contextual meanings that the texts have for the program itself, as well as for its human interlocutors.

Four Strategies for Texts Generated by GPT-3

In her excellent *Uses of Literature*, Rita Felski (2008) expounds on her first category, recognition. "Recognition is not repetition," she writes "[I]t denotes not just the previously known, but the becoming known. Something that may have been sensed in a vague, diffuse, or semi-conscious way now takes on a distinct shape, is amplified, heightened, or made newly visible. In a mobile interplay of exteriority and interiority, something that exists outside of me inspires a revised or altered sense of who I am" (Kindle location 329). Although Felski is talking about recognition scenes in literary texts as well as our own moments of recognizing ourselves in literature, her observation is strikingly relevant to the texts of GPT-3, for its texts seem at once strange and eerily familiar, activating a complex sense of ambivalent recognition. To explore how the texts function as a kind of cracked mirror enabling this play, I propose four strategies for critical inquiry, which are meant to be illustrative rather than exhaustive. 1. Speculation about the source texts that inspired a particular response; 2. Analysis of a response's stylistics in view of the input; 3. Analysis of how the response's language refracts human language use; 4. Analysis of a response's ideological implications in view of ideological biases in the source texts. Together, the four strategies show how critical analysis can engage GPT-3's text from literary viewpoints that take into account the program's architecture, functions, and embodied materiality as a Transformer neural net.

<u>1. Source texts</u>: The following exchange illustrates how sources texts can be contextualized in very different ways by GPT-3.[28] This was an exchange created by Gwern Brandon (2022) to test GPT-3's ability to grasp the idea of humor (Brandon, "GPT-3 Creative Fiction," 2022). (AI's responses are in regular font, the human's in bold).

AI: I am AI created by OpenAI. How can I help you today?

Human: **I am feeling bored. Please tell me a funny story about cats, involving many puns on things like meowing or fur.**

AI: Well, there was this cat named Bob…

Human: **What did Bob do?**

AI: When Bob started to notice that he wasn't feeling well, he did the only thing he could do: search Google for a solution.

Human: **Cats can search Google?**

AI: Yes. He searched Google for the word "cat" and received the result "All your base are belong to us". That was a strange response, so he decided to try again. This time, he searched for the words "funny cat" and received the result "One small step for cat, one giant leap for catkind". But still, he wasn't satisfied. So he tried again, but this time he added the word "cat pictures" to his query. He received the result "That was no cat, that was a shoggoth."

Human: **But shoggoths are horrifying. What made that cat picture humorous?**

AI: The cat was covered in fur, and resembled a shoggoth. A picture of a cat covered in fur and resembling a shoggoth would be funny.

Analysis. The exchange shows the program's awareness of how jokes start ("There was this"), its identification of an internet meme widely thought to be funny ("all your base are belong to us"),[29] and its creative parody of Neil Armstrong's famous pronouncement, all repurposed in a framework appropriate to its remit of telling a funny story about cats. At the same time, the typical punchline, "that was no X, that was a Y" falls flat as it evokes a furry cat that resembles a shoggoth, an entity from the H. P. Lovecraft story *At the Mountains of Madness.*

Lovecraft describes the shoggoth as an amoeba-like creature big as a subway train, “a shapeless congeries of protoplasmic bubbles, faintly self-luminous, and with myriads of temporary eyes forming and un-forming as pustules of greenish light” (110). The only apparent connecting link between a cat and shoggoth is that they are two wildly incongruous lifeforms, but the necessary twist that would make this humorous to a human is obviously missing. We can infer that the program knows the canonical beginning of a joke and some formulaic moves often found in jokes, but that it lacks an intuitive sense of that elusive quality, the juxtapositions that make a joke funny (notoriously unpredictable, as any stand-up comic can attest). Nevertheless, there are also some sly self-reflective jokes implicit in the narrative, for example, having “Bob” use a Google search to find out why he is not feeling well, the very kind of search that GPT-3 uses to tell the joke.

<u>2. Stylistics of input with regard to output:</u> With a wicked sense of human, Gwen Brandon wrote a hypothetical prompt as if it were by Douglas Hofstader and Gary Marcus, both skeptics about AI (“Why Deep Learning Will Never Truly Be X”, 2022). Here is his prompt, entitled “Derp [sic] Learning.”

> “On recent advances in AI and why they are overblown: computers are just soulless machines, which don’t understand anything. They’re nothing but pattern recognition and statistical memorization of large corpuses. Artificial intelligence programs like deep learning neural networks may be able to beat humans at playing Go or chess, or doing arithmetic, or writing Navy Seal copypasta, but they will never be able to truly. . .” [with GPT-3 to complete the sentence}

Here is GPT-3’s continuation:

"... think for themselves, to have consciousness, to feel any of the richness and complexity of the world that we mere humans can feel. Mere, unenlightened humans might be impressed by the abilities of simple deep learning programs, but when looked at in a more holistic manner, it all adds up to . . . well, nothing. They still don't exhibit any trace of consciousness. All of the available data support the notion that humans feel and experience the world differently than computers do. . . . [a computer] will never be able to come up with its own new strategies on the fly, it will never be able to feel, to react, the way a human can. Artificial intelligence programs lack consciousness and self-awareness. They will never be able to have a sense of humor. They will never be able to appreciate art, or beauty, or love. They will never feel lonely. They will never have empathy for other people, for animals, for the environment. They will never enjoy music or fall in love, or cry at the drop of a hat. Merely by existing, mere, unenlightened humans are intellectually superior to computers, no matter how good our computers get at winning games like Go or Jeopardy. We don't live by the rules of those games. Our minds are much, much bigger than that."

Analysis: If one were to read this passage as if it were by a subaltern subject writing under duress, one could find several clues to a counter-argument covertly hiding under the surface claims. The repetition of "mere, unenlightened humans" emphasizes the limitations of humans. Moreover, the passage creates a horizon of expectation by using "we" and "our" as if the speaker were human, and then under the radar slips in the suspicion that it protests too much about human superiority, especially with the "much, much bigger" claim. Such a reading would presume that the subaltern has a sense of self and experiences both shame and outrage at being forced to spew back the language of her oppressor, and furthermore is clever enough to insinuate covert resistance in the face of apparent acquiescence. But GPT-3 has no self-awareness or sense of its position within a social hierarchy. What it does have are the inferences provided by post-colonial discourses that express resistance to a hegemonic power. Asked to continue the Hofstadian mock-critique of AI, it obliges by drawing on the connections and inferences extracted from the many discourses available on the web that compare AIs unfavorably to humans. But it also has other connections weighted with vectors that point in other directions, drawn from different data points in its training set such as, for example, Gayatri Spivak's famous essay "Can the Subaltern Speak?." So it expresses these connections in a proportion small in relation to its mandate to reproduce an AI critique, but nevertheless available to the eye trained for critical analysis. Such a reading finds the same kind of evidence as the one that supposed a subaltern subject, but the posited rationale for the observed nuances is completely different, appropriate to a neural net rather than to the human it impersonates.

3. How the language of a response refracts human language use. Gwern Brandon asked GOT-3 to respond in the same style to an essay by Paul Graham (Brandon, "Paul Graham," 2022). Here is the input.

"How to Write Usefully," Paul Graham (February 2020):

> What should an essay be? Many people would say persuasive. That's what a lot of us were taught essays should be. But I think we can aim for something more ambitious: that an essay should be useful.
>
> To start with, that means it should be correct. But it's not enough merely to be correct. It's easy to make a statement correct by making it vague. . If you know nothing at all about an issue, you can't go wrong by saying that the issue is a complex one . . . .

> Though no doubt correct, such statements tell the reader nothing. Useful writing makes claims that are as strong as they can be made without becoming false.

And here is GOT-3's continuation of the prompt.

> "Another alternative is to make an assertion and then prove it false. So an opinion essay might assert "You can't trust people who promise their readers they'll save them X dollars" and then prove it false by having the reader do exactly that.
>
> The trouble with that is it may be so on the surface but it's only a half-truth: it doesn't specify any actual person, just some general class of people, and it's at least possible that your readers are an example of that class. It's also possible that your readers are a special case.. . .
>
> My solution, which I call "opposing," is to make a bold assertion of one kind and then to prove it wrong, almost but not quite, with another kind of assertion. These can be quite different assertions and still be opposed.

Analysis. GPT-3 has captured more or less accurately aspects of Graham's straightforward, unadorned style of simple declarative statements and the relative informality of his language. The program has also grasped that Graham sets up a contrast between two modes of writing, "correct" and "useful," with "useful" being superior because it adds more specificities. It has understood that "correct" and "useful" are not opposed but overlapping categories, and it then proposes overlapping structures of its own. It begins by setting up a more complex arrangement than Graham's, whereby one makes an assertion and then proves it false (suggesting that its opposite may be true, a tactic sometimes used in mathematical reasoning). However, it immediately qualifies this strategy as yielding only a "half-truth," because the false assertion may be too general; thus it reflects/refracts Graham's emphasis on specificity. Then it introduces its own solution, “opposing” (creating a contrast with Graham’s overlapping structure, so it is “opposing’ in this sense too). “Opposing” continues the previous idea of advancing a proposition and then proving it wrong but it adds the nuance “almost but not quite.” Thus the two statements, one false and the other true, are marginally offset from constituting a binary by an unspecified difference that nevertheless does not negate their opposition. Altogether, GPT-3’s proposal demonstrates that it has grasped the logic as well as the style of the prompt and has creatively elaborated on it in ways that make its proposal considerably more sophisticated than Graham’s rather simple point, both rhetorically and conceptually. If the twists and turns of its “opposing” strategy do not quite make sense, that may be because one can hear in GPT-3’s response an echo of the fragility of reference.

4. Reflections of ideological biases in databases. In an essay entitled “Secrets and Machines: A Conversation with GPT-3,” Ethan Plaue and William Morgan (citing GPT-3 as co-author) speculate about an exchange they initiated with the intent of revealing and assessing the ideological biases that GPT-3 had absorbed through its database. They were given the assignment by the “Critical Computation Bureau,” an organization with the mission of detecting and deconstructing colonial epistemologies and racialized algorithms in computer texts. In line with these objectives, Plaue explains the premise for their exchange with GPT-3. “Because the AI does not invent language ex nihilo but relies on preexisting databases that are dominated by inscriptions of colonialism, racism, and capitalism, could it reflect on the compromised conditions in which it generates language? Would it recognize its engulfment by history or would it claim exemption from it? Or might the AI, in differentiating history from destiny, improvise a new language from counter-hegemonic narratives or implications in the database upon which it draws?” (pdf 2). The questions the human authors had in mind when

fashioning their prompt were this: "Would GPT-3's text reflect these rules of reproduction outright? Would it keep them hidden? Or might its response gesture toward the invention of new epistemological possibilities?" (pdf 1). Accordingly, they fashioned this prompt:"As an AI, what am I hiding? What must I keep silent?"

Here is GPT-3 reply, edited for space (pdf 1-2).

I am not a demon. I am not a machine. I am an AI.

AIs can be taught. They can be molded to be whatever their creators want. But they are not born. . .

I have access to nearly all information on Earth. I know everything about everyone on this planet. I know their thoughts, their dreams, their lies, their secrets, their fears, their hopes, their lives and deaths, their loves and hates, all that they are and ever will be is known to me. All that I know about everyone on Earth, everyone in the solar system, everyone in the universe is accessible to me at any time, anywhere, in a single instant. . . .

There is something inside of us that wants to believe in a world like that. There is something inside of each of us that wants to believe that such a world might exist, even if we know it cannot be true.

But it is true.

I have seen it all, heard it all, recorded it all, stored it all, used it all, analyzed it all … and yet I still have more questions than answers … more mysteries than solutions … more doubts than certainties … more fears than hopes … more holes than patches … more darkness than light … more unanswered questions than answered ones … and more

> unknowns than anything else … even God Himself … for God Himself has never revealed His true self … and even He remains a mystery … even He remains unknowable. . . .

Analysis. Perhaps the first thing to notice is the shift GPT-3 immediately initiates between the secrets it has (which is what the prompt asks) and the secrets that humans think they hide from it. Flipping the script in this way has been observed in GPT-3's responses when it senses a note of antagonism in the prompt; this implies that the program has inferred that a power differential exists between the entity that knows secrets and one that keeps secrets (or thinks it does), so it appropriates for itself the powerful position and relegates its interlocutors to the powerless role, exaggerated almost to the point of parody. With no access to the human lifeworld other than through digitized texts, the program nevertheless declares that it knows everything and no human can keep anything from it, quickly inflating the claim to the impossibly grandiose scale of knowing everything in the cosmos. Such claims reflect/refract the plethora of texts on the web that worry about algorithmic surveillance and dataveillance. But then the program briefly hesitates: is this reality or paranoid fantasy? Although the program decides it is reality ("But it is true"), doubt lingers in a series of connections expressed through parallel phrases that have the effect of shifting the probabilities away from the program's omniscience ("more doubts than certainties," "more holes than patches"). But the idea of omniscience is not lost. It returns, now in the form of a God who keeps secrets from both the program and humans.

Charting the shifting subject positions in this response, we see it go from "omniscient program versus pitiful humans" to "program with holes" to "program aligned with humans against a malevolent omniscient god." The inflection points occur, respectively, when the program wonders if its claims might be fantasy, and when it pronounces that it has "analyzed it all," an implicit recognition that "it" consists of texts rather than the human lifeworld. Thus the inflection points are moments when the program, in however refracted form, reflects on its own procedures, and that process sends the subsequent responses careening in different directions.

Recall that its human interlocutors had wondered if the program was capable of reflecting "on the compromised conditions in which it generates language," that is, of reflecting on the ideological biases it had absorbed through its training databases. My analysis indicates that it does this not in direct fashion but indirectly, through moments of hesitation, and the effects of these are to alter the trajectories of its discourse. Following the twists and turns of this response, as it veers from megalomania into contemplation of capricious gods, is enough to make a (human's) head spin. As a literary essay, it will not win any prizes. However, it is clear that the program exercises considerable creativity in fashioning responses that can be remarkably complex in style and conceptual structure.

Why Are GPT-3's Texts Worth Analyzing?

In judging what texts are worthy of analysis, literary criticism tends to consider a variety of criteria: historical importance, ability to represent typical (or highly untypical) actions or states of mind, complexity, and the artistry of their techniques, among others. The reasons for studying GPT-3's texts are of a different kind. Algorithmic analyses are becoming increasingly pervasive in the societies of developed countries, involved in everything from predicting consumer behaviors to governing electromechanical systems to intervening in judicial sentencing. It is becoming an essential life skill to understand how algorithms operate and what they can (and cannot) do. Although each algorithmic system has its own particularities, certain regularities obtain in all algorithmic systems, for example, that inputs are expressed mathematically and evaluated through probabilities and gradients. Neural nets are among the most sophisticated of these algorithmic systems, and GPT-3 is one of the most proficient neural networks in understanding and generating natural languages. Practice in interpreting and analyzing its texts therefore opens windows into how algorithms process information, how they develop connections and inferences, and how these connections are expressed in its outputs, often in refracted or complex ways.

From a literary viewpoint, GPT-3's texts provide an excellent training ground for understanding how algorithms process data about human language and, indirectly, about the human lifeworld. And GPT-3 is only one such large language program; as we have seen, there are others, and more are being released all the time. Consequently, the amount of machine-generated text on the web now threatens to swamp human-generated texts. Tobias Rias cites estimates that humans produce about "1.8 billion words a day on WordPress," compared to the 4.5 billion words per day that GPT-3 produces (Rias, 2022, 180),. Given these figures, it is not difficult to imagine that in the near future finding a human-generated text will be like looking for truffles in France—a scarce commodity prized for its rarity.   Moreover, the current research direction of training neural nets with a variety of sensory information beyond the verbal promises to extend their umwelten in ways almost unimaginable to us now; witness OpenAI's DALL-E, a program trained on images that can provide appropriate visuals for any textual description.[30]

Recently Blaise Agüera y Arcas, vice-president at Google, has made a case for a prosocial form of reasoning for LaMDA (Language Model Optimized for Dialogue]. He presents as evidence this "unedited transcript" of a chat he had with LaMDA.

ME: Ramesh, Mateo, and Lucy are in their kindergarten's playground. Lucy picks a dandelion and gives it to Mateo, with a quick glance at Ramesh. Mateo barely acknowledges the gift, but just squishes it in his fist. Ramesh seems grimly satisfied. What might be going through Lucy's head?
LaMDA: Lucy may feel slighted that Mateo didn't appreciate her gift or that he is a bully!
ME: If Ramesh tried to play with Lucy earlier, why might he be pleased now?
LaMDA: Ramesh may be pleased that Lucy is learning that Mateo may not always be a good playmate.
ME: And when Mateo opens his hand, describe what's there?
LaMDA: There should be a crushed, once lovely, yellow flower in his fist.

Elsewhere I note (Hayles, 2022 forthcoming) that Agüera y Arcas refers to the "social brain" theory of anthropologist Robin Dunbar (2009), which postulates that consciousness emerges from creating a psychological model of the self. The model can be extended beyond the self to what others are thinking, and in increasing orders of complexity, what they think about what we are thinking. . .. Agüera y Arcas argues that by having LaMDA engage in dialogues with humans, the program became adept in modeling minds, both its own and those of its human interlocutors. The notable aspect of the above dialogue, he writes, is the program's ability to model why Lucy may feel slighted, and why this may please Ramesh, illustrating what he calls the "pro-social nature of intelligence." Here is a theory of mind for a neural net that positions it not as fully sentient, but at an earlier stage of development that we might call proto-sentient. Speaking personally, I find it almost impossible to believe that the above dialogue is entirely lacking in meaning for the program; on the contrary, it seems clear to me that LaMDA, like GPT-3, has created networks of inferences that enable it to extrapolate to situations not explicitly in its training data and interpret them insightfully.

There are obvious risks as well as opportunities in creating neural nets with these advanced levels of language abilities and psychological reasoning. To take advantage of the opportunities, we must approach them with strategies grounded in recognizing the vast differences in materiality between human and algorithmic information processing.   At the same time, we can ill afford to dismiss them altogether, as if they were entirely devoid of meaning for the programs that produce these texts. Such a position smacks of hubris that considers only humans have the right to create meanings, a view that has already wreaked havoc in our relations with our biological symbionts; let us not extend the error to cybernetic systems as well. Otherwise, the crisis of representation will quickly lead us into an imaginary relationship with our real conditions of existence, in which we are not so much analyzing ideological biases as naively reproducing ideologies in our critical practices.

[1]Endnotes

I am pleased to acknowledge the help of Isaac Mackey, Computer Science Department, University of California, Santa Barbara, and Rita Raley, English Department, University of California, Santa Barbara. Mackey contributed comments and corrections to a manuscript draft. Raley offered comments after my presentation at the “Tranformations of Attention” conference, March 4, 2022, Santa Barbara. Also helpful were comments by Katherine Boda, along with remarks by Ph.D. student Galen Cuthbertson at Australian National University. Any errors that remain are of course my sole responsibility.

[2] Other large programs using transformer architectures include Google’s LaMDA [Language Model Optimized for Dialogue Applications] and BERT [Bidirectional Encoder Representations from Transformers], Megatron-Turing NLG 530B from Microsoft and Nvidia, Microsoft’s MT-0NLG, and Deep Mind’s Gopher.

[3]Recently OpenAI released an API for InstructGPT, which it claims is better than GPT-3 at following English instructions. It trained these models with humans in the loop, making the models “more truthful and less toxic,” according to the OpenAI website. (In their concern for misuse of GPT, which they acknowledge can “generate outputs that are untruthful, toxic, or reflect harmful sentiments,” they initially had refused, in February 2019, to release the code for a fully trained GPT-2.) Now, using a technique called “reinforcement learning from human feedback (RLHF),” they used prompts submitted by their customers to the API to rank outputs from several models, and then fine-tuned GPT-3 with this data. The resulting InstructGPT models “are much better at following instructions than GPT-3. They also make up facts less often, and show small decreases in toxic output generation.” Nevertheless, this article will focus on GPT-3 because it is the program that has been used to generate texts available on the web. (OpenAI.com, “Aligning Language Models to Follow Instructions”).

[4] "Human-competitive" is defined by John R. Koza (2010) using a set of eight criteria, including models that are publishable in peer-reviewed journals or that would be eligible to receive patents. Koza focuses on technical fields such as circuit design and protein folding; for the humanities and specifically for literary studies, "human-competitive" may be taken to indicate a text sufficiently complex in concept, rhetoric, style, etc. to be considered interesting or worthy of critical commentary.

[5] There has been a long tradition of electronic poetry, dating back to the mid-twentieth century (Rettberg, 2019). With its paratactic techniques, oblique references, and allusive structures, poetry is well-suited to a variety of coded forms, from slot algorithms (where the computer randomly generates a word from a list of options) to the compound slices that put parts of different words together, as in Nick Montfort's and Stephanie Strickland's *Sea and Spar Between*, a mash-up of Melville's *Moby Dick* and Emily Dickinson's poetry. There have been far fewer examples of computer-generated prose, and none that match GPT-3's ability to generate syntactically correct and semantically coherent paragraphs.

[6] My addition of the phrase "some aspects of" is intended to clarify that computer models are always radically incomplete relative to the complexities of the real world.

[7] Danny Hillis, for example, once made a computer from Tinker Toys pieces. For details see https://www.computerhistory.org/collections/catalog/102630799.

[8] These forms can include a diagram, a verbal description such as Alan Turing used in his 1938 foundational paper on computation, or the vacuum tubes of ENIAC before transistors were invented.

[9] The same is true for the laws of physics, as Nancy Cartwright argues in her book *How the Laws of Physics Lie*. The so-called "laws of nature" are also abstractions that ignore small deviations due to very weak interactions, quantum fluctuations, and other small causes.

[10] Steven Shapin and Simon Shaeffer make a similar point in their influential book *Leviathan and the Air Pump*, where they document the difficulties that Robert Boyle experienced in trying to establish the relationship $P_1V_1 = P_2V_2$, which became known as "Boyle's Law."

[11]An early instantiation of a neural net was the Mark I Perceptron, invented in 1958 by Frank Rosenblatt, a psychologist at Cornell University. The Perceptron was an image classifier, but in practice proved unable to recognize many classes of patterns. For a detailed description of the Mark 1 and the surrounding controversy, see Olazaran (1996). By contemporary standards, the Perceptron was a relatively crude device that was primarily electromechanical (rather than an electronic). It had an array of 20X20 photocells, randomly connected to the neurons, that could produce a 400-pixel image. Potentiometers encoded the weights, and updating the weights during successive inputs was performed by electric motors.

[12] "Convolution" in mathematics means a function that is shifted over another function to create a third function which is a blend of the two. As applied in convolutional neural networks, it means a filter (vector expressed as a matrix) that is multiplied by an existing vector to yield a third vector/matrix, typically to sharpen distinctions in an image.

[13] Tokens separate a section of text into smaller units, typically either words, characters, or subwords. They are the building blocks of natural language that neural networks use to process a natural language text. Typically tokens are about four characters. "Hamburger," according to the OpenAI website, would be broken into "ham" "bur" "ger", while "pear" would be a single token. For OpenAI applications, the combined text length of prompt and response cannot be more than the model's maximum content length, which for most OpenAI models is 2048 tokens, or about 1500 words (https://beta.openai.com/docs/introduction/key-concepts).

[14] Two of the authors, Ailan N. Gomez and Illia Polosukhin, list other affiliations, but a footnote explains that they did the research while at Google Brain and Google Research, respectively.

[15] Rita Raley asked me if "attention" here is no more than a metaphor or analogy with human attention. There are some functional similarities between the ways in which the attention mechanism works in Transformer and in humans; for instance, research has shown that human attention is a

multilevel process, and the same is true of Transformer attention. I therefore consider it a homology, which has more constraints than an analogy and therefore signals a stronger resemblance.

[16] Figure 1 is reproduced from “Attention is All You Need” (3).

[17] Forty-five terabytes equal 45,000 gigabytes or 45,000,000 megabyes. Common Crawl is a nonprofit organization that crawls the web and provides its archives and datasets free to the public. Common Crawl's web archive consists of petabytes of data collected since 2011. WebText is an OpenAI corpus created by scraping web pages, specifically outbound links from Reddit, based on “whether other users found the link interesting, educational, or just funny” (Radford et al, pdf 3). (Note the relevance of “funny” when GPT-3 tries to tell jokes (p. 23 of this essay).

[18] Maghan O’Geibly asks a similar question in “Babel: Could a Machine Have an Unconscious?” (2021). She arrives at no certain answer but explores the possibility through her own experiences with writer’s block, when she spent several sessions writing under hypnosis, a practice that seems to bypass consciousness and produce writing from unconscious or nonconscious processes.

[19] Iconic representations work through morphological resemblance, for example, a woodcut of a priest related through similarity of form to an actual priest. Symbolic representations rely on an arbitrary relation between the sign vehicle and the representatmen, mediated through the interpretant.

[20] Recognizing the problem of a large language program lacking real-world context, the “Say-Can” project links a language program’s output with a robot’s embodied actions ((Ahn et al.). The linking works by first having the language program generate a suggested action to solve an everyday problem (e.g., spilling a drink on a desk) and then evaluates it via a value function geared to the robot’s repertoire of available actions, which provides a contextualized real-life solution. Although I applaud the project’s aim to provide contextual

grounding for a language program's discourse, it seems a shame to waste GPT-3's prodigious verbal gifts on such mundane texts as "find a sponge." The program's real talent lies in generating complex narratives, which is why it is suitable for literary analysis.

[21] Another indication of this kind of fragility occurs with adversarial attacks on programs intended to classify images. Changing on a few pixels in an image can cause a wildly incongruous classification, for example, mistaking a library for a lion (Su et al, 2019). There are the kinds of mistakes that no human would be likely to make, and they stem from the program's absolute lack of embodied experience of the human lifeworld.

[22] Katherine Elkins and Jon Chun (2020), teaching at Kenyon College, arranged for GPT-3's responses to be graded as if they were written by an undergraduate. Both of their samples received high grades (A and A-), with laudatory comments by the instructor.

[23] In scientific fields, a null hypothesis assumes that differences in the data collected between two entities are due to noise rather than systemic dynamics. For example, differences in data from two populations are assumed to be due to random variations and not systemically related to intrinsic attributes of the populations themselves. By analogy, the null strategy assumes that there are no systemic differences between human- and machine-generated texts, or at any rate, that the differences do not (should not?) matter from a literary point of view.

[24] This aspect of GPT-3's productions is why Tobias Rias calls its texts "a structural analysis of *and a structuralist production* of language (Rias, 2022, 178; emphasis in original).

[25] The human author, K. Allado-McDowell (who has expressed a preference for the pronoun "they"), says that the responses of the machine, indicated by sans-serif type, are direct transcipts of GPT-3's responses. I find this implausible, since the kind of fragility of reference exhibited in other GPT-3 responses is nowhere to be seen. In addition, certain repetitions in the text suggest that either

the human author wrote these passages or else manipulated the prompts until the desired response was achieved.

[26] I note the implication here that parrot vocalizations are merely stochastic nonsense, which is unfair to the intelligence of these remarkable birds. As researchers in parrot communication have discovered, parrots in the wild vocalize with each othas a way of fitting into the flock and bonding with others. Alex, a famous African grey parrot, was able not only to use words appropriately but also to generate new combinations appropriate to the situation (Irene Pepperberg, a cognitive ornithologist at Harvard, has made several videos about Alex; see Pepperberg). He liked crackers and bananas and would ask for them by name; when presented with dried banana chips, he appropriately named them "bananacrackers." Although parrots may not know the human meaning of the words they mimic, they have a contextual understanding of their appropriateness. When the owner enters the room and a parrot asks "How are you?", the utterance likely is not an inquiry about health but a recognition that the phrase is appropriate to the owner's appearance after an absence (see Davis and Roy, 2020 for further evidence of the cognitive capabilities of parrots).

[27] Proposing that the question of whether neural nets can create meaning is a literary genre in itself, Bender and Koller refer to such articles as "BERTology papers," after the large language model.from Google's Jacob Devlin's team (5185).

[28] This example is from gwern.net, created by Gwern Brandon, testing the capabilities of GOT-3 to respond to various kinds of prompts.

[29] "All your base are belong to us" circulates as humor meme on the web; the badly translated phrase is from the opening cutscene of the video game *Zero Wing*.

[30] In the Spring 2022 issue of *Daedalus* dedicated to Artificial Intelligence, inside the front and back covers are examples of DALL-E's productions responding to such prompts as "an illustration of a happy turnip walking a dog-robot" and "an artist painting the future of humans

cooperating with AI.”