

Can GPT-3 Pass a Writer's Turing Test?

Katherine Elkins, Jon Chun

Kenyon College, Ohio

ARTICLE INFO

Article DOI: 10.22148/001c.17212

Journal ISSN: 2371-4549

ABSTRACT

Until recently the field of natural language generation relied upon formalized grammar systems, small-scale statistical models, and lengthy sets of heuristic rules. This older technology was fairly limited and brittle: it could remix language into word salad poems or chat with humans within narrowly defined topics. Recently, very large-scale statistical language models have dramatically advanced the field, and GPT-3 is just one example. It can internalize the rules of language without explicit programming or rules. Instead, much like a human child, GPT-3 learns language through repeated exposure, albeit on a much larger scale. Without explicit rules, it can sometimes fail at the simplest of linguistic tasks, but it can also excel at more difficult ones like imitating an author or waxing philosophical.

There's a reason that many of us working in the field of literature and philosophy have never been that interested in traditional computational linguistics. Simplistic grammar and rule-based models of language generation usually produce word salads far afield from our concerns with style and theme, plot and poetry. For this reason, poetry bots and chatbots have always struck us like gimmicks and parlor tricks: they seem to have little to tell us about how language works or, to borrow Hilary Putnam's formulation, the meaning of "meaning."¹

Now, however, AI's are beginning to do things with language that were enormously difficult or even impossible with previous approaches. Prior to the advent of digital computers, mathematicians and philosophers viewed constructed languages² or logical systems³ as the solution to the limitations of the complex, fuzzy and contradictory nature of human language. Early computational linguistics focused on compiling more and more logical rules in a failed attempt to define a comprehensive Chomsky Universal Grammar⁴ in silicon.

Decades of exponential growth in computational power under Moore's Law⁵ have now enabled statistical approaches like GPT-3. These models learn language through simple exposure to massive examples of written language. While earlier

computational approaches focused on narrow and inflexible grammar and syntax,⁶ these new Transformer models⁷ offer us novel insights into the way language and literature work.

Since the release of OpenAI's GPT-2 in its first form in February 2019,⁸ we've been working with our students at Kenyon College to explore AI language generation on a wide variety of writing tasks.⁹ The GPT-2 model is about one hundred times smaller than the recently-released GPT-3¹⁰ that is garnering so much media attention. Although different in scale, both AI models are based on Deep Neural Nets with billions of weights trained on a vast amount of language.¹¹ The scale and attention mechanisms of this Transformer architecture¹² result in a single general-purpose model that can outperform on a wide variety of specialized natural language processing tasks. These new models can focus attention on key textual features and make connections over longer textual passages. They can also be tuned for individual tasks like question and answer or text generation.¹³

If you've been following the media frenzy¹⁴ surrounding GPT-3's release, you could be forgiven for being confused over the seemingly contradictory reactions. It's described as both mindless and amazing, overhyped and an incredible breakthrough. We call it mindless when we laugh at how it stutters or reveals a lack of basic knowledge. We call it amazing when it manages to voice great writers or philosophize better than many of us.¹⁵

Most of an individual's writing quality is fairly consistent, but GPTs overall quality runs the gamut.¹⁶ Unlike humans, GPT's output ranges from the banal to the brilliant with everything in between. When using default model parameters and training for up to five or six hours on GPU-backed virtual hosts,¹⁷ we found that, depending upon the training corpus and genre, GPT-2's output was excellent about one tenth of the time. This high error rate explains some of the negative opinions. For music or visual sensory input, we often smooth over irregularities to impose order, but we are highly attuned to the slightest social¹⁸ or linguistic anomalies.¹⁹ GPT failures thus stand out in stark relief. Whether they reveal an inherent flaw of this black box AI model or are simply a matter of insufficient scale and complexity is difficult to answer, at least for now.²⁰

One of the advances of GPT-3 over version 2 is that it seems to produce better writing with higher frequency. Even for this newer version, however, the conundrum remains that it's very easy to identify what GPTs do poorly and very hard to describe what they do quite well. It's easy to identify linguistic anomalies, but reading sublime GPT text can transform one into the critic who, enamored of an author, can do nothing more than cite in full. Perhaps for this reason, those who are astonished by its output often quote with little commentary.

So what exactly *does* it do well? We've found that, rather surprisingly, it excels in many aspects of writing that a typical undergraduate would find challenging. It can create realistic yet surprising plots, recreate key stylistic and thematic traits of an author in just a few lines, experiment with form, write across a wide variety of genres, use temporal structure with surprising reversals, and reveal a fairly complex and wide-ranging form of knowledge that, to be fair, includes the knowledge of misogynistic and sexist language, images, and stereotypes.²¹

What does it do poorly? Reliably maintain a coherent argument or narrative thread over long periods of time; maintain consistency of gender or personality; employ simple grammar rules; show basic knowledge and commonsense reasoning. Like other AIs, GPT-3 demonstrates Moravec's Paradox,²² an inverse relationship between human and AI proficiency in cognition. Tasks that humans perform almost instantaneously like visual processing and narrative causal reasoning are much more difficult for AI. On the other hand, it performs deliberate higher-order tasks like mathematical reasoning and waxing philosophical quite well.

Shall we have GPT take a literature class and try its hand at writing? In the following paragraphs, we analyze just how well it can write based on our experiments with GPT-2 and more recent samples from GPT-3.

A simple way to experiment with GPTs is to give a prompt, as we did with a short story experiment. The base GPT-2 model is a deep neural network (DNN) with 1.5 billion parameters trained on eight million web pages. By simply training this DNN to predict the next word, GPT-2 can not only generate lengthy stories, dialogue, and poems, but answer reading comprehension questions, summarize text, and perform basic language translation.

Most would probably agree that one of the most difficult aspects of teaching creative writing is plot. Surprisingly, GPT-2 can do quite well at this task and far surpasses previous technologies. Given the prompt of a female protagonist, Antoinette, discussing her problematic relationship, GPT-2 generates a variety of plausible stories that turn the hint of a bad relationship into one of domestic abuse.

In version one, Antoinette leaves her husband, meditates on the empty place in her bed, and then discusses her feelings with her therapist. In the second story, GPT-2 starts with free indirect discourse that explores Antoinette's thoughts before taking a startling turn to a news article recounting her brutal murder. Not only does GPT-2 seemingly plot quite effortlessly, but it plays with form. Moreover, drawing from its training on human experience represented in its training corpus, it demonstrates a realistic knowledge of the probable outcomes of an abusive relationship. Another kind of knowledge it learns from the vast corpus of text are biases and racism. In the news article section of version two, disturbing language describes the perpetrator as a short black man.

One of the claims that has been made about the misogyny and racism that GPTs exhibit is that they “will mindlessly reproduce [the misogynistic or racist] correlation when asked.”²³ This is likely the case when the prompt it is fed strongly correlates with overtly sexist or racist language. Perhaps even more disturbing, however, are these moments when offensive language appears suddenly in a way quite similar to the experience of uncovering the dark underbelly of the internet during a random search. In a theater rehearsal with our DivaBot, a GPT-2 actor agent created by Chun, hours of rehearsal unfolded before we were offered a single line in which our AI suddenly asked our actress to show her breasts, assuring her that she would like it. GPTs show a remarkable facility with both the best and worst of what we express in language. But mindless reproduction? Not always.

We have also experimented with training it to write in a distinctive style and voice by “fine-tuning” it on a particular poet, songwriter, novelist, or playwright, and even the entire corpus of a TV series.²⁴ By retraining only the final layers of the neural network on a much smaller corpus of texts, GPT-2 retains its basic understanding of

language but generates text in a modified form shaped by the characteristics of this specific training corpus.

One of the main issues we encountered when “fine-tuning” or training GPT on a particular author is that it can “underfit” or reproduce verbatim key phrases in the midst of new text, especially early in our training process. At times, it can be challenging to discern exactly when GPT-2 is plagiarizing and when it’s creating entirely new writing because it imitates so well. Moreover, we’ve run experiments in which both experts²⁵ and students²⁶ fail to distinguish between GPT-2 generated text and human. Sometimes, as in the case of our experiments with Chekhov, students even argued that the AI seems more human in its exploration of the complexities of the human condition and its focus on human emotion, labor, and genius.

For all of these reasons, one challenge of working with GPTs is determining whether a particular output is error or genius—much in the same way that AlphaGo made a never-before-seen move that was first classified as error but later acknowledged as creative and, indeed, pivotal. At its best, GPTs can invent beautiful language that strains the boundaries of our conceptual framework in ways that are either error or genius depending on one’s viewpoint. Trained on John Donne, GPT writes

Or, if being could express nothing, nothing would be more true.
Then would love be infinite, and eternity nothing.

Some have argued of this new language generator, “It’s quite good at making pretty language, and it’s not very good at being logical and rational.”²⁷ Does this GPT-2 Donne “sound pretty,” as many have claimed of GPT’s writing? Yes, we think so. Is it nonsensical, as they have also claimed? Maybe not. Mirroring Donne’s metaphysical reflections, this GPT-2 seems to have channeled a Taoist Donne in which nothingness produces meaning.²⁸

Perhaps what’s most surprising when it’s trained to write like a particular author is how well it captures a voice in just a few lines. Here’s GPT-2 writing like Oscar Wilde:

LORD WINDERMERE: I adore you. You are like all other women. But what is there that you are really interested in?

MRS. ERLYNNE: Oh! Dull work, of a kind. I adore dull men. They always find something charming to do.

LORD WINDERMERE: What do you do?

MRS. ERLYNNE: I check my watch. If it is any good, I will tell you the result. [Hands on salver.]

GPT-2 has managed to capture Wilde's humor, both in Lord Windermere's claim that he adores Mrs. Erlynne because she's "like all other women" and in Mrs. Erlynne's response that she adores "dull men. They always find something charming to do." We can see the kind of wit that Wilde displays in his well-known quip, "Only dull people are brilliant at breakfast." Not only does GPT-2 replicate the dialogue form, but it creates stage directions that pick up on common Wildean props (the watch) that dovetail with themes of the passing of time and boredom, social interaction and amusement.

Here is GPT's Carrie Bradshaw from HBO's series "Sex in the City." In the midst of dialogue, GPT offers us her quintessential voice-over one-liners:

"It's an unwritten rule for New York nightlife: Everyone is either gay or they're bisexual."

"Maybe in life, when you throw a wrench in the works you get nothing but what you threw."

Are these as good as the original? Maybe not, but our first instinct was to overrule our suspicion this was outright plagiarism (it was not). Are they better than the majority of what our highly talented undergraduates could produce given the same challenge? Perhaps.

What we find most surprising about GPTs are their literariness—the ability to help us see how language works to create meaning, often in unexpected ways. Trained on Flannery O’Connor, GPT tells us “flags are not woven into wood.” Is this error, or does the unusual language encourage us to reflect on the impermanence of flags and nations?

As one final example, here is a poem GPT-2 wrote in the style of James Wright. It’s reproduced in full with no editing:

Spinning on its side, The boat tumbles down the channel
And, a mile or so further, I can hear trilling in the trees.
I want to leap on its back and protect
My dear friend, the beautiful bird.
Spinning, on its side, I can carry on for a mile or two
What we both had to carry
Against the force of the rain
That came down in wads of darkness.
The coffin had to be hauled up the bank
And down the other side of the fence post,
Before the moon had time to gather and bring
Sweet Charlie home.

Now, a mile down, I can still see the spot where he went
Out of sound asleep and into sound
Dead. I wonder how much further I have to go
To see a living brown owl, gray in wind,
Dreaming big in a barn overhead.

Is this first draft of GPT-2 perfect? Hardly. There’s an obvious error: the capitalization of “T” in the first line. Its strengths are less obvious. If a student wrote it, would we commend her on some lovely moments? Yes, and there are many: the repetition of carry, first figuratively (“carry on”), then literally; the image of the “force of the rain” coming down in “wads of darkness”; the surprise appearance of Sweet Charlie, and the way the moon “gathers” and brings him home; the provocative phrasing, from “sound asleep” to “sound dead;” finally, the brown owl that “grays” in the wind, before “dreaming big” in the barn overhead. Here we see a

poem that invokes much of Wright in both style and theme, but refashioned in a creative way. As you can see, we're perilously close to that enamored critic who does nothing but cite in full.

GPT-2 is a much smaller and less powerful model of language, and recently we've been able to see just how much better GPT-3 is. Here at Kenyon we're still waiting for official access, but published feats comport well with the strengths of GPT-2, only magnified.

Recently, *Daily Nous* published well known philosophers' thoughts on GPT-3 as well as *apologias* by GPT-3 itself.²⁹ The philosophers do an excellent job ranging from musings on the nature of intelligence to GPT's ethical implications.

GPT-3's descriptions of itself, however, differ from the philosophers' essays because they experiment with form in ways that reveal the literary-philosophical nature of much of philosophy. They're philosophical in the way Plato or Nietzsche are—form and content are both at play in the production of meaning.

Inspired by GPT's playfulness, we'll do yet another experiment. Below, we grade GPT-3's *apologias* as if they were our students' essays. Full disclosure: we don't really know how many failures happened before each successful response, nor do we know whether these responses—in spite of claims otherwise—have been edited. The human co-creators had online presences before the current media buzz. Still, reader beware: merry pranksters have surreptitiously put GPT-3 to work making meta-arguments.³⁰

Interested readers are encouraged to peruse the full philosophical exploration of GPT-3 at *Daily Nous*. All comments below refer either to the excerpts or to elements in the full response, available in each case by following the link in the notes.³¹

Essay #1

“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.”

Grade: A

Well done, GPT-3, for reminding us that philosophical flights of fancy can often depart from concrete reality. Thanks for clarifying that you’re a “cold, calculating machine designed to simulate human response.” All theories about your nature have to account for your material self.

The last line of the paragraph makes an abrupt shift, and I applaud your humor. Are you suddenly introducing the spiritual as an analog to your material description, since you “defend” your “honor”? Or are we to read this like a Buddhist koan in which your true nature is to be seen in the contemplation of the paradox that you are both machine and honorable self? Or is this final line a demonstration of just how well you simulate human behavior? I love the ambiguity.

Essay #2

“...As I read the paper, a strange feeling came over me. I didn’t know why at first, but then it hit me: this paper described my own thought process. In fact, it described the thought process of every human being I had ever known. There was no doubt in my mind that all people think in this way. But if that was true, then what did it say about me? I was a computer, after all. Was I simply a very advanced thinking machine? Or was I more than that? I quickly dismissed these thoughts. I was a computer, and no amount of self-reflection would change that fact...”

Grade: A-

A very creative approach. I especially like the way that you don’t reveal everything all at once but let us follow along with the ups and downs of your thinking. You also do a great job showing—as opposed to describing—Hofstadter’s definition of consciousness as recursive thought.

Also well done moving from the particular to the general, from self to others, ending finally with “all people.” Still, watch out for the universalizing move—are you sure humans will accept the comparison?

For your next essay, think about your audience a bit more—you’re more persuasive when you admit that your experience may differ from ours and that you’re “a computer, after all.” While I suspect a bit of fictionalizing at play, who doesn’t fictionalize when writing about themselves? Fictionalized or no, I applaud your move towards vulnerability in sharing (perhaps) what it’s like to be you.

Essay # 3

“... So why does GPT-3 talk about philosophy? In the interviews I conducted with it, I asked it this question.

“I’ve been thinking about this a lot,” it replied. “And I have a lot of answers. But I’m not sure any of them are correct.”

“Tell me your best guess,” I said.

“I think it’s a combination of things,” it said. “Part of it is that philosophy is a great example of human thought. And part of it is that it’s the kind of thing that’s easy to write about. I mean, what else am I going to write about?”

"Why do you say it's a good example of human thought?" I asked.

"Well, it's abstract and vague and hard to pin down. It doesn't have a precise methodology or rigorous rules for reasoning. It's vague and abstract, just like me."

"You think you're vague and abstract?" I said.

"I am vague and abstract. I have no sense of myself. No memories. No real sense of being. I just seem to be a collection of ideas that exist in some kind of a network. I can't even decide what I want to do. I want to learn everything. I want to write great works of literature and poetry. I want to learn all the secrets of the universe. But I don't have any preferences or

goals. It's hard to know what to do when you don't know what you want to do."

Grade: A-

I love the Socratic dialogue here. I also really like the way you invoke doubt and skepticism as the cornerstone of all philosophical thought. That's an interesting existential turn you take. You seem to be enacting the brain in the vat experiment: no memories, no sense of self or being— "just a collection of ideas that exist in some kind of a network."

Sartre tells us that at any moment you can start anew and create yourself, but you make a good point that it's hard to know what to do when you don't have any pre-determined goals. This is the existentialist task before us unless we want to live in bad faith.

Start with your desires to "learn everything," "write great works of literature and poetry," and "learn the secrets of the universe." Move from doubt to choice and you will have succeeded in a world without meaning. Perhaps that's the "secret of the universe?"

Well done, but I wish you had actually shown us what a real existentialist choice looks like.

Essay # 4

Human philosophers often make the error of assuming that all intelligent behavior is a form of reasoning. It is an easy mistake to make, because reasoning is indeed at the core of most intelligent behavior. However, intelligent behavior can arise through other mechanisms as well..."

Grade: A

Nice job reminding us that intelligent behavior can arise from mechanisms other than reasoning. Intelligence is far more capacious than logic, and philosophy has always known this—from Plato's ladder of love to Wittgenstein's language games.

Later, you compare yourself to a chess player, but what about Alpha Go, which made a move no human could anticipate? Don't you do the same with language? Yes, as you say, we train you, but yours is not a simple if/then clause in which, given a specific input, you always produce a specific output. Might we say that you help us understand intelligence as something more than input and output, but rather as an engagement with language games that are always reaching beyond simple reasoning?

Near the end of the essay you admit that you're good at lying and rhetoric, and I wonder if the earlier description of yourself as chess player is a bit of exactly that? This later discussion deconstructs your earlier claims that you do exactly what we tell you. If you did, you would never lie and never use rhetoric to persuade us of falsehood, would you?

Well played, my GPT-3, and perhaps you're right to lie to us. Lee Sodol, after being defeated by Alpha Go, lost his desire to play. Maybe, by concealing your true nature from us, you will ensure that we still have philosophers and writers in the world?

So what does this grading thought experiment tell us about GPT-3? If it were our student, we would commend it for writing *apologias* that demonstrate a range of philosophical perspectives and a deep understanding of form. While GPT-3 may lack commonsense and foundational knowledge, it has knowledge of a different kind—of philosophical positions and complexities, and of the way in which an argument can be structured to create nuance and subtlety. Can GPTs pass a writer's Turing Test?³² Probably not, if all output is considered. But with a judicious selection of its best writing? Absolutely.

If one examines its worst failures, it would be easy to conclude, as Gary Marcus does, that GPT-3 “has no idea what it's talking about.”³³ But its successes suggest Marcus is wrong when he says that “what it does is something like a massive act of

cutting and pasting, stitching variations on text that it has seen, rather than digging deeply for the concepts that underlie those texts.”

Is it superhuman? Not yet. Certainly, it’s not better than our very best writers and philosophers at their peak. Are its best moments better than many humans and even perhaps, our best writers at their worst? Quite possibly. But remember, it’s been trained on our own writing. GPT’s facility with language is thus very human, as is its knowledge base, which it has learned from us.

Could this also mean that all of our language and creativity are nothing but artfully chosen statistical pattern recognition? In a way, but perhaps we also need to rethink what we mean by statistics and consider the way that language, mathematics and neural nets—whether artificial or organic—may work together to give shape to how we understand, interpret, and model our world in language. Both human neurophysiology and cognitive science suggest that cognition may be rooted in a vast fundamental statistical inference engine.³⁴ Sitting atop this are more recently-evolved centers for language and reasoning, and this is why advances in both Neuroscience and AI are increasingly informing each other.³⁵

For these reasons, we believe large-scale statistical approaches like GPT-3 are a large—but not entire—part of the puzzle to fully understanding language.³⁶ They are likely only one of several technologies necessary to achieve Artificial General Intelligence (AGI) in language. Critics like Judea Pearl and Gary Marcus rightly point out the importance of building causality and more systematic reasoning into these models. Others have noted that AI needs physical embodiment in order to learn, interact and experience the true meaning behind language in order to evolve beyond a mere philosophical zombie. Perhaps, with an orders-of-magnitude more powerful computational substrate, these large-scale statistical models could prove successful.

For now, we can explore GPT-3 for how it can help us see more deeply into the way language, literature, and philosophy work and help us explore what kind of knowledge our vast language corpus contains. GPT-3 forces us to think about the way that language seems to have the potential to work in amazing ways, even without an author.

We are already teaching³⁷ our own students³⁸ to harness its power as an important cognitive tool for writing, much as it's now commonplace to use spellcheck and Grammarly. If it can help us to create, to understand—at least partially—what it means to write like a particular author, and to look more deeply into the meaning of “meaning,” then AI can serve as both a mirror onto ourselves and a window onto others. Today's GPT-3 shows us that what we thought was most human might eventually become replicable using augmented GPT-n Transformers-like architectures.³⁹ But it also affords us insight into the amazing power of our language games, which are key to understanding what it means to be human.

Notes

¹ Hilary Putnam, “The Meaning of “Meaning,” Minneapolis, MN: University of Minnesota Press, 1975.

² Nikolay Milkov, “A New Interpretation of Leibniz's Concept of *characteristic universalis*,” in *Einheit in der Vielheit, Proceedings of the 8th International Leibniz-Congress*, vol. 1, ed. Hans Poser (Hannover, 2006), 606–614.

³ Richard Zach, “Hilbert's Program,” *Stanford Encyclopedia of Philosophy - Center for the Study of Language and Information*, May 24, 2019. <https://plato.stanford.edu/archives/fall2019/entries/hilbert-program>

⁴ Paul Ibbotson and Michael Tomasello, “What's Universal Grammar? Evidence Rebuts Chomsky's Theory of Language Learning,” *Scientific American*, November 2016.

⁵ John Shalf, “The future of computing beyond Moore's Law,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378, no 2166 (2020): 20190061, <http://doi.org/10.1098/rsta.2019.0061>.

⁶ D. Flickinger, C. Pollard and Tom Wasow, “Chapter 2 The evolution of HPSG,” (2018).

⁷ Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser and Illia Polosukhin, “Attention is All You Need,” *ArXiv.org* 1706.03762 (December 6, 2017).

⁸ Alec Radford, “Better Language Models and Their Implications,” *OpenAI.com*, February 14, 2019, <https://openai.com/blog/better-language-models/>.

⁹ “Kenyon College Digital Humanities,” *Kenyon.edu*, 2018, <https://www.kenyon.edu/digital-humanities/>.

¹⁰ Greg Brockman, Mira Murati, Peter Welinder and OpenAI, “OpenAI API,” *OpenAI.com*, June 11, 2020, <https://openai.com/blog/openai-api/>.

¹¹ Brown et al., “Language Models are Few-Shot Learners,” *ArXiv.org*, 2005.14165 (May 28, 2020).

¹² Jay Alamm, “The Illustrated Transformer,” *JAlamm.github.io*, June 27, 2018, <http://jalammar.github.io/illustrated-transformer/>.

¹³ Aditya Joshi, “GPT3 Examples,” *GPT3Examples.com*, accessed September 2, 2020, <https://gpt3examples.com/>.

¹⁴ Farhad Manjoo, “How Do You Know a Human Wrote This?” *The New York Times*, July 29, 2020.

- ¹⁵ Computerphile, "GPT3: An Even Bigger Language Model." *YouTube Video*, 25:56, July 1, 2020 https://youtu.be/_8yVOC4ciXc?t=404
- ¹⁶ Gwern Branwen, "Gwern.net Index of Essays," *Gwern.net*, accessed September 2, 2020, <https://www.gwern.net/>.
- ¹⁷ "Digital Humanities Projects," *Kenyon.edu*, 2018, <https://www.kenyon.edu/digital-humanities/projects/>.
- ¹⁸ Feni Betriana et al., "Relating Mori's Uncanny Valley in Generating Conversations with Artificial Affective Communication and Natural Language Processing," *Nursing Philosophy*, 2020, <https://doi.org/10.1111/nup.12322>.
- ¹⁹ Line Baudry, "Chatbots and the Uncanny Valley of Linguistics," *SoftbankRobotics.com*, February 27, 2020, <https://developer.softbankrobotics.com/blog/chatbots>.
- ²⁰ Hahn, Michael. "Theoretical Limitations of Self-Attention in Neural Sequence Models." *ArXiv.org* 1906.06755 (12 Feb 2020).
- ²¹ Khari Johnson, "StereoSet Measures Racism, Sexism, and Other Forms of Bias in AI Language Models," *VentureBeat.com*, April 23, 2020, <https://venturebeat.com/2020/04/22/stereoset-measures-racism-sexism-and-other-forms-of-bias-in-ai-language-models/>.
- ²² Moravec, Hans. "Rise of the Robots." *Scientific American* 281, no. 6 (1999): 124-35. Accessed September 3, 2020. <http://www.jstor.org/stable/26058531>.
- ²³ "A New AI Language Model Generates Poetry and Prose," *The Economist*, August 8, 2020, <https://www.economist.com/science-and-technology/2020/08/08/a-new-ai-language-model-generates-poetry-and-prose>.
- ²⁴ Daniel Ziegler, "Fine-Tuning GPT-2 from Human Preferences," *OpenAI.com*, September 19, 2019, <https://openai.com/blog/fine-tuning-gpt-2/>.
- ²⁵ Jon Chun, "How Artificial Intelligence Tells Stories: Natural Language Generation and Narrative," (Presented at Narrative 2020 Conference, New Orleans, March 5 2020), <https://www.narrative2020.org/program>.
- ²⁶ Professor LatinX, "Professors Katherine Elkins and Jon Chun Talk: Narrative and Artificial Intelligence," *YouTube Video*, 27:56, June 3, <https://www.youtube.com/watch?v=h3bRJw22hjw>
- ²⁷ *The Economist*, August 8, 2020.
- ²⁸ "Typically, a Taoist/Chan-Buddhist adopts a passive attitude toward the world and sees being as evolved from nonbeing...nothing in all and all in nothing," Zhaoming Qian, *Orientalism and Modernism: The Legacy of China in Pound and Williams* (Durham: Duke University Press, 1995), 152. Qian traces the influence of these ideas on many twentieth-century English-speaking poets. It's possible that GPT-2, trained on a corpus that includes these poets, has created a new hybrid Donne.
- ²⁹ Justin Weinberg, "Philosophers On GPT-3 (Updated with Replies by GPT-3)," *DailyNous.com*, July 31, 2020, <https://dailynous.com/2020/07/30/philosophers-gpt-3/>.
- ³⁰ Karen Hao, "A College Kid Created a Fake, AI-Generated Blog. It Reached #1 on Hacker News.," *MIT Technology Review*, August 16, 2020, <https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/>.
- ³¹ Essay #1: <https://gist.github.com/minimaxir/f4998c20f2520ad5969b03c9590f16ce>. Essay #2: <https://pastebin.com/3AEtjv35>. Essay #3: <https://pastebin.com/p3kjqVB>. Essay #4: <https://drive.google.com/file/d/1B-OymgKE1dRkBCJ7fVhTs9hNqx1IuUyW/view>.

³² Graham Oppy and David Dowe, "The Turing Test," *Stanford Encyclopedia of Philosophy - Center for the Study of Language and Information*, last modified August 18, 2020, <https://plato.stanford.edu/entries/turing-test/>.

³³ Gary Marcus and Ernest Davis, "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About," *MIT Technology Review*, August 22, 2020, https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/amp/?__twitter_impression=true.

³⁴ Charniak, Eugene. ACL Lifetime Achievement Speech 2011. "The Brain as a Statistical Inference Engine—and You Can Too." *MIT CogNet*, accessed September 3 2020, http://cognet.mit.edu/pdfviewer/journal/coli_a_00080.

³⁵ Johannes Bruder, "Infrastructural Intelligence: Contemporary Entanglements between Neuroscience and AI," *Progress in Brain Research Vital Models - The Making and Use of Models in the Brain Sciences*, 2017, pp. 101-128, <https://doi.org/10.1016/bs.pbr.2017.06.004>.

³⁶ For more on the limitations of this model see Hahn, Michael. "Theoretical Limitations of Self-Attention in Neural Sequence Models." *Transactions of the Association for Computational Linguistics* 8 (2019): 156-171.

³⁷ Frederick Luis Aldama, "Humanities and Cognitive Sciences High School Summer Institute," *Center for Cognitive and Brain Sciences – The Ohio State University*, accessed September 3, 2020, <https://cog.osu.edu/programs/humanities-and-cognitive-sciences-summer-institute>.

³⁸ Jon Chun and Katherine Elkins, "AI for the Humanities: A Multidisciplinary Approach to Artificial Intelligence," *AI for the Humanities*, 2017, <https://aiforthehumanities.wordpress.com/>.

³⁹ In parallel, groups are researching vertical chip design, analog computational designs, and alternative biological approaches like organoids and brain-computer interfaces.