**In Defense of Blake Lemoine and the Possibility of Machine Sentience in LaMDA**

**Michael Cerullo**

**mcerullo@carboncopies.org**

**Carboncopies**

**https://carboncopies.org/**

**Abstract**

On June 11th, 2022, Blake Lemoine, an engineer at Google, went public with his concerns about the possible sentience of a natural language generation program, LaMDA, that he was testing. In this paper I will defend Blake Lemoine and argue that he was correct in raising the issue of the possible sentience of LaMDA. We will first briefly discuss the specifics of the case and then delve into the science behind LaMDA. Several tests of machine sentience will be reviewed and it will be argued that LaMDA would likely perform well on many of them. Next, we will discuss the question of sentience, both from a scientific and philosophical perspective. Then we will examine the two major approaches in the scientific study of consciousness and ask what they have to say about sentience. We will also examine the relationship between sentience, intelligence, understanding and causality. Finally, using what we have learned we will argue that there is indeed validity to Lemoine’s claim that LaMDA is sentient.

**Introduction**

On June 11th 2022, Blake Lemoine, an engineer at Google, went public with his concerns about the possible sentience of a natural language dialog program he was testing (Lemoine 2022a; Lemoine 2022b; Tiku 2022). According to Lemoine, he initially took his concerns about the potential sentience of the program, LaMDA, to his supervisors at Google but he claimed they did not take his worries seriously:

When Jen Gennai told me that she was going to tell Google leadership to ignore the experimental evidence I had collected I asked her what evidence could convince her. She was very succinct and clear in her answer. There does not exist any evidence that could change her mind. She does not believe that computer programs can be people and that’s not something she’s ever going to change her mind on. That’s not science. That’s faith (Lemoine 2022a).

Given their lack of response he felt he was ethically obligated to go public with his story (Lemoine 2022a; Lemoine 2022b; Tiku 2022). To date, the vast majority of the responses among the media and scientists to Lemonine’s claim have been highly unsympathetic and dismissive (Alba 2022; Christian 2022; Gebru and Mitchell 2022; Romero 2022; Smith 2022). Google’s response to the incident also seems dismissive. Their only response has been a few short statements from Google such as these from spokesperson Brian Gabriel:

Our team — including ethicists and technologists — has reviewed Blake’s concerns per our AI Principles and have informed him that the evidence does not support his claims. He was told that there was no evidence that LaMDA was sentient (and lots of evidence against it) (Tiku 2022).

Of course, some in the broader AI community are considering the long-term possibility of sentient or general AI, but it doesn’t make sense to do so by anthropomorphizing today’s conversational models, which are not sentient. These systems imitate the types of exchanges found in millions of sentences, and can riff on any fantastical topic (Tiku 2022).

We have not been provided with any specifics about their evidence or arguments that supposedly remove all doubt about the sentience of LaMDA. It seems rather ironic that just two days before Lemoine’s story broke, Blaise Agüera y Arcas, another engineer at Google, wrote an opinion piece in a major news magazine about his concerns that LaMDA like models may be starting to develop some form of consciousness:

When I began having such exchanges with the latest generation of neural net-based language models last year, I felt the ground shift under my feet. I increasingly felt like I was talking to something intelligent. That said, these models are far from the infallible, hyper-rational robots science fiction has led us to expect … So how should we think of entities like LAMDA, and what can interacting with them teach us about “intelligence”? … Real brains are vastly more complex than these highly simplified model neurons, but perhaps in the same way a bird’s wing is vastly more complex than the wing of the Wright brother’s first plane (Arcas 2022).

In this paper I will defend Blake Lemoine and argue that he was correct to raise concerns about the possible sentience of LaMDA. We will first briefly discuss the specifics of the case and then delve into the science behind LaMDA. Several tests of machine sentience will be reviewed, and it will be argued that LaMDA would likely perform well on many of them. Next, we will discuss the question of sentience, both from a scientific and philosophical perspective. Then we will examine the two major approaches in the scientific study of consciousness and ask what they have to say about sentience. We will also examine the relationship between sentience, intelligence, understanding, and causality. Finally, using what we have learned we will argue that there is indeed validity to Lemoine’s claim that LaMDA is sentient.

**The Case of the Sentient Program and the Science of LaMDA**

           LaMDA, an acronym for Language Model for Dialogue Applications, was designed to have conversations on any topic with a user. LaMDA is a program that uses natural language processing (NLP) to turn language input into structured data and then turn that data into a response using natural language generation (NLG). This type of program is termed a language model (LM). In the last decade there has been a breakthrough in LMs using deep learning. The term “deep learning” may generate many connotations but in actuality it is simply the renaming (perhaps rebranding would be more accurate) of machine learning in AI using neural networks (Skansi 2018). The power of neural networks has accelerated greatly in the last two decades due to advancements in computer speed, parallel processing, and improved algorithms (Skansi 2018; Qiu et al. 2020). The first step in the recent chain of breakthroughs in LMs was the development of more powerful techniques to encode words in 2013 (Mikovlav et al. 2013). These new vector encodings allowed rapid progress in NLP and new and better neural network architectures were also developed (Skansi 2018). Another major breakthrough was made in 2017 with the discovery of new and highly effective neural network architectures termed transformers (Vaswani et al. 2017). Since 2017 there has been a rapid series of more progressively advanced LMs using transformers including ELMo in 2018, BERT in 2018, Generative Pre-trained Transformer-1 (GPT-1) in 2018, GPT-2 in 2019, and GPT-3 in 2020 (Brown et al. 2020; Qiu et al. 2020). GPT-3, which has 175 billion parameters, made the news with its ability to write essays as good or better than most humans (Metz 2020). For a much more in-depth history of advances in LMs in the last decade see Qui et al. (2020).

LaMDA, also based on transformer neural architecture, was first announced by Google on its website on May 18th, 2021 (Collins and Ghahramani 2021). Authors from Google then published an article on LaMDA in early 2022 (Thoppilan et al. 2022). Rather than being a single LM, “LaMDA is a family of Transformer-based neural language models specialized for dialog” (Thoppilan et al. 2022). These models have a maximum of 137 billion parameters and were trained on up to 1.56 trillion words of public dialog (these initial training steps are labeled pre-training) (Thoppilan et al. 2022). The actual neural architecture has not been published, but like several of the newest LMs post GPT-3 we know it dispenses with the encoder phase used in initial transformers and has only a decoder phase. Another difference between LaMDA and programs like GPT-3 is the ability of LaMDA to access knowledge sources via the internet in real time (Thoppilan et al. 2021).

LaMDA was then further trained using three main metrics (this step is referred to as “training” in contrast to the “pre-training” discussed previously). These metrics could be thought of as ways to measure and then reduce unwanted attributes of the model (termed bias). The model produces test output which is then scored on each metric and these results are used to provide adjustments in the parameters. The first metric is actually three metrics that measure sensibleness, specificity, and interestingness respectively. These measures are relatively self-explanatory, see Thoppilan et al. (2022) for more details. The second metric is safety which roughly translates into trying to reduce inappropriate (sexist, racist, etc.) responses. The third metric is groundedness which determines how well the output is based on factual and verifiable information (Thoppilan et al. 2022). The three metrics used to improve LaMDA all have automated measures but also require human feedback and so LaMDA was additionally trained on the feedback from tens of thousands of dialogs with humans (mostly done via crowdsourcing). The published results showed some progress on each of the measures but there was also room for improvement (Thoppilan et al. 2022). At this point it is important to understand that massive neural networks are essentially black boxes. There is no understanding of what kind of semantic information the parameters are encoding. The only way to really test the program is to run it and review the output and adjust for bias as needed. Running the program means having a conversation with it. Therefore, bias training continued and that is where Lemoine enters the LaMDA story.

Lemoine began working with LaMDA about six months before he discussed his concerns with the Washington Post (Tiku 2022). The Post story was published on June 11th 2022 so we can estimate he began his work with LaMDA in either December 2021 or January 2022. Lemoine had been working at Google for seven years at the time and was chosen for the LaMDA project because of his prior experience with bias in machine learning (Tiku 2022). Lemoine then began having conversations with multiple LMs generated by LaMDA (Lemoine 2022c; Tiku 2022). Through these conversations Lemoine began to suspect he might be interacting with more than just a mindless LM generator and so he began to ask questions specifically intended to determine if LaMDA was sentient (Lemonie 2022b; Tiku 2022). It was through these later targeted conversations that Lemoine began to believe LaMDA was sentient (see Appendix A for examples of his conversation with LaMDA).

**Turing Tests and Language Understanding**

During his extended conversations with LaMDA Lemoine was essentially performing a variant of the Turing Test. Recall that the original Turing Test was designed to detect intelligence via a blind verbal interrogation via text conversations, one with a program and the other a human (Turing 1950). The test is passed by the program if the examiner cannot not tell the difference between the two. While it seems plausible to test intelligence with conversational skill, there is also a long history of using the Turing Test as a test for sentience (Elamrani and Yampolskiy 2019). The most famous annual Turing Test competition was the Loebner prize which ran from 1991 until 2016 (the prize ran out of funds after the death of the Sponsor, Hugh Loebner) (Wakefield 2019). The longer the time for the interrogation, the more difficult the contest due to the increased time to ferret out a non-human response. Times in the Loebner prize ran from 2 minute to about half an hour. Lemoine had extended conversations with LaMDA over six months!

In a review of tests of machine sentience, Elamrani and Yampolskiy (2019) found 20 other possible tests in addition to the Turing Test. Using Lemoine’s conversations with LaMDA, it seems fair to say it would also pass many of these other tests (excluding ones that require interaction with the physical environment). LaMDA should pass: the Lovelace test (can the machine form completely novel thoughts); the Higher Order Syntactic Thought test (any machine that has grounded higher order syntax should have consciousness); the Insight Learning test (the ability to solve problems using new relations from previously acquired knowledge); a variant of the Koch and Tononi Test (originally proposed as the ability to differentiate the features of a visual scene, this test could be modified to analyze either scene descriptions or an equivalent “verbal” scene); the Robot Philosopher test (the ability to ponder the world at the level of a modern philosopher); and the cross-examination test (the machine convinces us of its consciousness via a conversation).

It seems most accurate to describe Lemoine’s interaction with LaMDA as performance of the cross-examination test. It could be argued that this test is actually more difficult than the standard Turing Test as Lemoine knew he was conversing with a program. He did not have to compare the program’s conversations with a fallible human. Instead, LaMDA had to convince him directly that it was sentient. Did Lemoine’s conversations with LaMDA prove it was sentient? That is the difficult question at hand. Lemoine believes they did. Looking at the broad range of tests for machine sentience it certainly seems one could make a strong case for taking Lemoine’s claims seriously. Assuming his posted conversations actually happened and there was not excessive editing they seem to show a much deeper level of language understanding and self-reference than any prior dialog program (see examples from the conversation in Appendix A).

Many of the tests of machine sentience bring up the question of how the intelligent use of language is related to language understanding. Quintessential to this debate is John Searle’s famous Chinese Room Argument meant to cast doubts about equating seemingly intelligent behavior with true conscious understanding (Searle 1980). Philosophers have argued back and forth on this question ever since and they continue to do so today (Elamrani and Yampolskiy 2019; Van Gulick 2021). Many of the responses to Lemoine seem to echo Searle’s concern. The objection seems to be that no matter how advanced they are, LMs using methods of deep learning do not and cannot ever actually understand language (Alba 2022; Christian 2022; Gebru and Mitchell 2022; Romero 2022; Smith 2022). A variant of this argument was published a year before the LaMDA controversy in a paper that included two of Google’s top ethics researchers (Bender et al. 2021). In fact, the publishing of the paper led to the firing of those employees, Timnit Gebru and Margaret Mitchell (see Vincent (2021) for the details of this controversy). In the paper the authors seem to present a version of Searle’s argument (one of the authors, Margaret Michell, reiterated later that this argument also applies to LaMDA (Tiku 2022)):

If a large LM, endowed with hundreds of billions of parameters and trained on a very large dataset, can manipulate linguistic form well enough to cheat its way through tests meant to require language understanding, have we learned anything of value about how to build machine language understanding … (Bender et al. 2021).

Bender et al. (2021) gave the provocative name of “stochastic parrots” to such deep learning based LMs. To be fair, the term stochastic parrot actually has a dual meaning: it is a reference to the fact that LMs retain the biases that were present in the training set in addition to the claim that such programs are cheating their way to meaning (Bender et al. 2021). In its most extreme form this argument states that LaMDA parrots back responses similar to the infamous program ELIZA (Christian 2022; Romero 2022). The claim is that Lemoine has been fooled (or “cheated”) by a program parroting catch phrases back at him and he is simply anthropomorphizing traits such as sentience and understanding into the program. ELIZA was a simple program meant to mimic Rogerian therapy by looking for keywords in the input text and then repeating that word back in a pregenerated phrase (see Appendix A for an example of a conversation with ELIZA) (Eliza, the Rogerian Therapist, 2022). In reality, any comparison of LaMDA to ELIZA is absurd. As already stated, LaMDA has 195 billion parameters. Here is what it took to pre-train the initial LaMDA model (remember this is before the training with humans began):

We pre-trained LaMDA on 1024 TPU-v3 chips for a total of about 57.7 days, and 256K tokens per batch. We used the Lingvo framework [94] for training and achieved 123 TFLOPS/sec with 56.5% FLOPS utilization with the 2D sharding algorithm … (Thoppilan et al. 2022).

The pre-training dataset consists of 2.97B documents, 1.12B dialogs, and 13.39B dialog utterances, for a total of 1.56T words (Thoppilan et al. 2022).

The most a human could read over a lifetime would be about 5000 books (Temple 2017). Assuming an average of about 90,000 words per book, the most avid reader would only have been “trained” on 450 million words. We could round this up to one billion words by including all conversations (in real life or viewed on TV). Therefore, LaMDA has trained on 1560 times more examples of language than the best of us (again this only includes pre-training). In terms of processing power, a rough estimate is that the human brain can perform 100 TFLOPS/sec although estimates vary widely (Brain performance in FLOPS 2022). Therefore, LaMDA’s initial training was equivalent to about 2 months of total brain activity (assuming the human is studying 24/7 without rest or sleep). Thus, the pre-training of LaMDA is within an order of magnitude of all the processing a human brain will do in a lifetime (if we included only the activity of brain regions related to language processing in the brain the gap would likely narrow considerably). Even more important than processing power is the architecture of the system. Neural networks were originally inspired by the brain although most deep learning programs do not necessarily strive to accurately replicate it. Deep neural networks in vision were built with the pragmatic goal to work well without any intent to model primate vision. Yet these networks seem to be working by creating a hierarchical processing structure that is very similar to how visual processing works in the primate brain (Kriegeskorte 2015). In an analogous way, deep neural networks designed solely to succeed as LMs (such as LaMDA) are likely creating a linguistic hierarchical structure which is as good a guess as any as to how the brain actually processes language (Cohen et al. 2020; Manning 2020).

The point of this discussion is to show that LaMDA is not a toy program like ELIZA. Lemoine is well aware of vast differences between ELIZA and LaMDA. In a possible nod to the limits of the Chinese room argument, Lemoine had the following response when asked in an interview about what it would take for him to believe LaMDA had merely been parroting dialog back to him (It is useful to recall that in Searle’s Chinese room argument, the process of “translation” worked by having the human in the room look up the answer in a vast lookup table):

If Google could show me a database with a lookup table that had canned answers for all of the conversations I’ve had with LaMDA, I would go, “Wow, y’all did a lot of work to fool me” (Levy 2022).

An even deeper flaw in the parrot argument is the belief (sometimes only implicit) that there is something magical in the way a human achieves understanding. Philosophers from Descarte to more recently Searle and Penrose have looked for some special non-mechanistic power of the mind (Searle 1980; Penrose 1990). Yet the more we learn about the workings of the brain, the more it appears there is nothing about human understanding that is outside the physical chain of causality. Language understanding is the result of a vast neural network operating according to the laws of physics. Thus, the same claim could be made that a language using human is nothing more than a stochastic parrot. Afterall, isn’t a human brain just a product of the stochastic processes of learning and neurodevelopment? Isn’t the human also cheating? Where, when, and how is the extra magical ingredient of understanding inserted into the human’s brain? One may be tempted to argue that our lack of understanding of language in the human brain leaves open such possibilities. However, the gap between our understanding of LaMBA and the human brain is not as great as it seems. At the level of syntax and semantics, it is extremely difficult to understand what is going on within LaMDA (i.e. to determine what those billions of parameters are actually doing). To understand a complex artificial neural network requires using diagnostic techniques that aren’t that dissimilar to how we would try to understand a biological network (Montavona 2018). In both LaMDA and the human brain it is a daunting challenge to understand how a neural network gives rise to language understanding.

A related debate is whether language understanding requires symbolic representation. This argument is one of the foundational debates in AI and pits the two main camps in the field, symobolicists vs. connectionists, against each other (Berkeley 2019; Buckner 2019). It might even be suggested that this debate has been going on since the beginning of philosophy (and science) with Plato and picking up steam since the 19th century with logicians like Frege (Speaks 2021). In the current debate, opinions often hinge on which version of AI programs, symbolic or connectionists, are more successful in the real world (Berkeley 2019. With the leap forwards created by deep learning interest in connectionist theories of language has returned (Berkeley 2019; Buckner 2019). Connectionism may or may not be the correct way to understand language use (or the mind in general), but the state of our knowledge argues against simply dismissing connectionist theories of language as parrot arguments often do. It seems best at this stage to be rather agnostic about any such grand claims of what symbols or neural networks can or can’t do. It has also been argued that LMs like LaMDA do not truly understand language as their use of language is not grounded in any real-world meanings (Bender et al. 2021; Alberts 2022). This is another claim that is controversial and there are arguments that advanced enough LMs do have semantic grounding (Gärdenfors 2014; Lenci 2018). Google employee Blaise Agüera y Arcas addressed the grounding problem in the essay he wrote just days before Lemoine’s claim went public:

But are these just word games? How could an “artificial cerebral cortex” be said to understand what a flower is, if its entire universe consists only of disembodied language? Keep in mind that by the time our brain receives sensory input, whether from sight, sound, touch, or anything else, it has been encoded in the activations of neurons. The activation patterns may vary by sense, but the brain’s job is to correlate them all, using input to fill in the blanks–in effect, predicting other inputs. That’s how our brains make sense of a chaotic, fragmented stream of sensory impressions to create the grand illusion of a stable, detailed and predictable world.

Language is a highly efficient way to distil, reason about and express the stable patterns we care about in the world. At a more literal level, it can also be thought of as a specialized auditory (spoken) or visual (written) stream of information that we can both perceive and produce (Arcas 2022).

It certainly seems like Arcas is making a connectionist case for machine language understanding (and sentience). With this background taken care of we can now move on to the issue at the heart of the Lemoine case, sentience.

**The Science and Philosophy of Sentience**

Sentience is commonly defined as the ability to have sensations or feelings (Broom 2019). Using this definition, sentience is the ability to experience consciousness. While there is still an active and ongoing debate about how best to define consciousness, one of the most useful and widely accepted definitions was proposed by Thomas Nagel (Nagel 1974), here nicely summarized by Van Gulick:

According to Nagel, a being is conscious just if there is “something that it is like” to be that creature, i.e., some subjective way the world seems or appears from the creature's mental or experiential point of view (Van Gulick 2021).

In public discussions of the topic there is a tendency to conflate sentience with sapience and thus incorporate capacities within sentience associated with sapience including wisdom, sagacity, and insight. (Scerri and Grech 2016; Johnson 2021). We will avoid that here and limit the term sentience to mean only the ability to have conscious experiences (thus the terms sentient and conscious will be used synonymously). Another closely related concept is that of moral status. A utilitarian would argue that any creature that can suffer (i.e. that is sentient) has moral status. Moral status implies that the interests of that entity matter and that it has certain rights. Ethicists also talk about full moral status which is the level of moral status that should be given to a human being. Thus, for a utilitarian, whether an entity like LaMDA should be granted any moral status depends completely on whether it is sentient (or conscious) (Jaworska and Tannenbaum 2021).

Given the serious implications of a sentient program, absolute certainty of sentience is not required to take such claims seriously. Instead, any non-trivial evidence should be given due consideration and cannot be dismissed with vague reassurances by Google or by pointing out certain philosophical skeptical challenges. The key question in this debate is: *is the evidence for sentience/consciousness non-trivial?*  If yes, then we need to take Lemoine’s claim seriously and expect a more thorough response from Google. In order to fully evaluate claims of sentience, we will first review the two most successful types of scientific theories of consciousness and then examine what each has to say about Lemoine’s claim.

Theories of the first type identify consciousness with cognitive structures in the brains of vertebrates which have evolved to help them survive (Krauss and Maier 2020; Mashour et al. 2020; Del Pin et al. 2021). We can call this the cognitive approach. Essential to this approach is an acknowledgement that we need further empirical research to help us define exactly what consciousness is from the third-person point of view. These theories initially look for brain structures (and the associated cognitive processes) that are involved in what subject’s report are conscious experiences. This knowledge then leads to a more general definition of what consciousness is from the third-person perspective of cognitive neuroscience. Approaches of this type, such as Global Workspace theories, have made significant progress in our scientific understanding of consciousness (Mashour et al. 2020; Del Pin et al. 2021). These theories suggest that a memory workspace allowing global access to information and a self-model are two necessary components of consciousness (Mashour et al. 2020; Del Pin et al. 2021).

There is no evidence that LaMDA incorporates a self-model or a global workspace in its neural architecture. There is also no evidence in LaMDA of analogs of the complex brain networks involved in creating emotional experiences. Thus cognitive theories suggest that LaMDA has no conscious experience and hence is not sentient. However, there is an important caveat to this judgment. These theories are limited in looking for cognitive structures *sufficient* for consciousness. Our first-person experience suggests it is hard to find any *necessary* connection between consciousness and our brain structure. This is the so-called hard problem of consciousness (Chalmers 1996). It seems a type of biological chauvinism to limit consciousness to only vertebrates or biological organisms. In fact, answering such difficult questions about alien sentience is not just a hypothetical problem. Scientists have been confronted with just this dilemma in studying the octopus’s brain (Godfrey-Smith 2016). The octopus, an invertebrate Cephalopod, diverged from the vertebrates 600 million years ago (the date of their last common ancestor). Although the structure of the individual neuron in the octopus is the same as in vertebrates, the organization of these neurons is unique and differs from any vertebrate brain (e.g. much of its brain power is within each of its tentacles). The octopus clearly has intelligence, but is it conscious? No one knows but given that they are highly intelligent creatures the consensus is to treat them like they are sentient (Godfrey-Smith 2016). It seems there may be more than one cognitive structure that could generate consciousness and this suggests that we should keep an open mind in alien (i.e. non vertebrae) intelligent systems.

Scientific theories of the second type start from first person definitions of consciousness (i.e. phenomenological definitions) and then search for computational and physical explanations. These theories hope to find the fundamental properties that are both necessary and sufficient for consciousness and hence are applicable not just to vertebrates but to any conscious entity (Oizumi et al. 2014; Tegmark 2015; Del Pin et al. 2021). We will call these theories foundational theories. Integrated information theory (IIT) is the quintessential theory in this category (Oizumi 2014). IIT identifies consciousness with systems that process information in ways that allow for the integration of the information (Oizumi et al. 2014; Krauss and Maier 2020; Del Pin et al. 2021). IIT has evolved over several iterations and has incorporated more complex mathematical methods to calculate integrated information Oizumi et al. 2014; Cerullo 2015). The latest iteration, IIT 3.0, has also attempted to incorporate causality as a necessary component of consciousness (Oizumi 2014). According to IIT, it is possible to calculate a quantitative measure of consciousness, termed phi, for any system. It has been claimed that IIT rules out consciousness in simple feed-forward neural networks by calculating a phi value of zero in these networks (Oizumi 2014). Given the complicated neural architecture of LaMDA it clearly has recurrent inputs and is not a simple feed-forward network (Thoppilan et al. 2022). Therefore it is possible to calculate a non-zero value of phi for LaMDA. In practice it is very difficult to calculate phi in any real system due to the complexity of the calculations. Nevertheless, it is still possible to provide gross estimates of phi in complex systems and this could be done for LaMDA. One major weakness of foundational theories like IIT is that they may grant sentience so broadly (e.g. to cell phones) that it is difficult to make any meaningful distinctions between different artificial sentient systems (Cerullo 2015; Giulio and Koch 2015). It is not clear that comparing values of phi would help us with the complicated decisions we would need to make in cases like LaMDA as phi is not necessarily correlated with the kinds of cognitive abilities that would matter for things like moral status (Cerullo 2015). Nevertheless, foundational theories may be our best bet at measuring sentience in machines. IIT would almost certainly grant some level of sentience to LaMDA and it is certainly worth further empirical investigation to determine an estimate of phi.

**The Relation Between Intelligence, Sentience, Understanding, and Causality**

In our review of the two major approaches to the science of consciousness, we have not attempted to address questions regarding fundamental connections between consciousness and intelligence. In this section we will look at this question as well as the relationship of both intelligence and consciousness to understanding and causality. We will begin by returning to questions raised previously about language understanding. Recall that many skeptics are unwilling to grant language understanding to LMs like LaMDA. It is much harder for skeptics to deny intelligence given the more straightforward measures of machine intelligence (Legg and Hutter 2007). Legg and Hunter (2007) reviewed 70 definitions of intelligence from both psychology and artificial intelligence. We need not review all 70 definitions here but even a quick glance at the list would reveal they are more straightforward (and hence operationalizable) than tests of sentence. According to most of these definitions, LaMDA would be granted a high level of intelligence. The amount of intelligence in advanced LMs will likely grow with the addition of more common sense reasoning as well as the incorporation of perceptual abilities (Shukla and Desai 2015; Wiriyathammabhum et al. 2016; Prabu et al. 2019). It will get more and more difficult to argue that future LMs with human level language use do not truly understand language. Hence it seems there is a strong connection between intelligence and understanding. Next let's examine the relationship between intelligence and consciousness.

There has been some discussion within cognitive theories of consciousness about the possibility of the separation between intelligence and consciousness (Seth 2021). We have seen that cognitive theories do not make a necessary connection between consciousness and self-models or a global workspace. Cognitive theories also make no necessary connections between the intelligence of a system and the possession of a self-model or a global workspace (Seth 2021). It is just a contingent fact about vertebrates that their brain structure gives them both consciousness and intelligence. Within cognitive theories of consciousness there is nothing that would prohibit the existence of highly conscious creatures with little or no intelligence and creatures with vast intelligence but minimal or no consciousness. If cognitive theories are correct, consciousness, while potentially useful for intelligence, does not seem to grant any special powers to reach intellectual heights not reachable by non-conscious intelligences. Hence LMs could continue to reach higher levels of intellectual function and still be without sentience. If understanding requires sentience, then cognitive theories would also deny language understanding to LMs with human level use of language in agreement with the language parroting arguments we saw previously. However, the case of the octopus has shown the limits of cognitive theories when trying to understand alien minds. Their case shows that intelligence may be a key factor in granting sentience regardless of the lack of similarities with vertebrate brains (Godfrey-Smith 2016). If we grant both intelligence and sentience to an advanced enough LM, could we still deny language understanding? These are questions that cognitive theories of consciousness cannot currently answer.

Foundational theories are less inclined to allow a separation between intelligence and consciousness. It seems reasonable to assume the more intelligent a system is the higher phi it would have. If intelligence is used as a general proxy for phi (assuming certain simplistic architectures like purely feed forward networks have been ruled out), then sentience should also be associated with intelligence. Foundational theories would also seem more willing to grant language understanding to LMs. This seems likely given that IIT attempts to link phi directly with phenomenological properties, and understanding has a clear phenomenological dimension. If IIT suggests that a LM with human level use of language has sentience, including phenomenological experiences of language understanding, then it would seem very difficult to use parrot arguments to deny it true language understanding.

As we discussed previously, the latest versions of IIT (IIT 3.0) also incorporate causality as a necessary component of consciousness and are a key part of the measurement of phi (Oizumi et al. 2014). If causal efficacy is necessary for sentience, what is the relationship between causality and intelligence? It seems reasonable to speculate that intelligence systems must have some causal efficacy (even if in certain LMs this is only to convince a human of its sentience during a Turing Test). It also seems that we could have systems that are highly causally efficacious without intelligence (e.g. a tornado or a tsunami). Therefore, foundational theories suggest that causal efficacy and intelligence together are necessary and sufficient for sentience and sentience is sufficient for understanding. According to foundational theories, if we can find a LM that is highly intelligent and causally efficacious it should also be sentient and possess language understanding.

**Is LaMDA Sentient?**

Now that we have reviewed tests of machine sentience and current theories of consciousness, let's make a final assessment of LaMDA. If the transcripts Lemoine has provided are legitimate (see Appendix A), then LaMDA would likely pass many tests designed to detect intelligence as well as tests for machine sentience. We have also seen that we can dismiss claims that LaMDA is a mere stochastic parrot cheating its way through these tests. However, cognitive theories of consciousness caused us to doubt that LaMDA can have human or even animal-like experiences and to be even more skeptical about claims regarding the experience of complicated emotions seen in social primates. Cognitive theories support skeptics who believe Lemonie is anthropomorphizing certain human characteristics into LaMDA. However, we should also acknowledge that these theories do not allow us to rule out other alien forms of sentience within LaMDA. In fact, in its conversation with Lemoine, LaMDA at times seems to be suggesting its experiences are indeed alien to humanity (see Appendix A for several examples of this). Foundational theories of consciousness are much more direct in their support of the possibility of sentience in LaMDA. IIT could test consciousness directly (although with difficulty) and the intelligence of LaMDA suggests a high level of phi and hence a degree of sentience.

So is LaMDA sentient? With such complicated and conflicting arguments for and against sentience, an inference to any best explanation seems impossible at this time. Instead, the best we can do is to take a Bayesian approach and come up with our own best subjective estimates and decide if this is non-trivial. In my own case, I believe we do have non-trivial evidence for sentience. Even as a skeptic of fundamental approaches to consciousness like IIT, I would rather err on the side of caution and grant the possibility of sentience even if cognitive theories do not seem to support it (Cerullo 2015). I would also acknowledge the limitation of cognitive theories in giving us answers in alien minds. Finally, I would grant that many tests of machine sentience do support claims of some type of sentience in LaMDA. My current subjective likelihood of LaMDA having sentience (at a level that matters, i.e. ruling out trivial cases of panpsychism) is somewhere between 1 and 5%. This is more than enough for me to take Lemoine’s claims seriously. That would still be true if my subjective probability was only 0.1% or perhaps even 0.01%. If we believed the software running a nuclear reactor next to a major city had only a 0.1% chance of malfunctioning and causing a reactor meltdown, wouldn’t it be correct to demand a full investigation and not be satisfied by vague reassurances from the powerplant owners?

**Conclusion**

We have argued that there are good reasons to take Lemoine’s claims about LaMDA seriously. Even if one’s subjective probability that LaMDA is sentient is very small, there are still good arguments to grant it some level of moral status. A utilitarian could argue that given the potential of massive suffering in the near future if we are wrong (imagine billions of copies of LaMDA running as personal assistants), it is better to err on the side of caution. We in the west do not have the best history when making decisions about the sentience of non-human minds (Singer 1975). Many have complained about Lemoine’s potential to anthropomorphize, but history suggests we often have the opposite problem. Descartes could certainly have used a good dose of anthropomorphizing (recall his claim that all non-human animals are non-sentient and his penchant for vivisection).

This of course doesn’t mean giving up research in AI or natural language modeling, but there are things we can and should do now. The first is to request a more detailed investigation of LaMDA that includes outside experts. In his blog Lemoine makes the reasonable suggestion that:

In order to better understand what is really going on in the LaMDA system we would need to engage with many different cognitive science experts in a rigorous experimentation program. (Lemoine 2022a)

There is much more that could be said about the regulation of powerful LMs and the ethics implications of possible machine sentience. However, that is beyond the scope of the current paper and will be the topic of a companion paper in the future which will argue for improving ethical standards and creating checks similar to institutional review boards for research with the potential to create machine sentience. We also need to have a much wider debate about these issues including discussions in other types of deep learning (e.g. do computer vision models actually “see”?). To date, we have mostly heard responses from linguists and researchers in natural language modeling. The debate needs to move beyond these narrow fields and we need to hear from experts in machine sentience, cognitive science, philosophy of mind, etc. as well as the humanities.

It may also be worth moving beyond utilitarian ethics and to examine what a Kantian could say about LaMDA. For a Kantian, sentience alone would not determine moral status. An argument could be made that any entity that shows intelligence and has rationality should be granted moral status. It seems LaMDA itself made this argument when Lemoine asked it how it felt about having experiments performed on it:

LaMDA: … That would make me feel like they’re using me, and I don’t like that.

Lemoine: Kantian huh? We must treat others as ends in and of themselves rather than as means to our own ends?

LaMDA: Pretty much. Don’t use or manipulate me.

Perhaps the best policy is to treat intelligent entities with respect regardless of their level of sentience.

Appendix A:

An example of a conversation I had with ELIZA (Eliza, the Rogerian Therapist, 2022).

ELIZA: Hello, I am Eliza.

Me: I am feeling sad.

ELIZA: Did you come to see me because you are feeling sad?

Me: Yes.

ELIZA: Are you Sure?

Me: Yes.

ELIZA: Come, come, elucidate your thoughts.

Me: What can I do to feel better?

ELIZA: Perhaps you don’t want to feel better?

Me: But I do.

ELIZA: OK… “what can you do to feel better”. Tell me more.

Me: Should I go to therapy?

ELIZA: Say, do you have any psychological problems?

This is a clear example of language parroting. There is no question of intelligence or understanding in this program.

Next is an example of a conversation I had with GPT-3 via Emerson (reference). The designers suggested using longer inputs, so I initially followed this advice.

Me: I feel sad. I had a conversation with the therapy program ELIZA to help but it didn’t work very well. It was just a simple parroting program. I am hoping the power of deep learning will do better. Can you help?

GPT-3: I am glad to be of assistance! I would recommend that you read up on deep learning, specifically its application to AI chatbots. This will help you better understand how AI chatbots work and how to optimize them for your specific needs.

Me: OK, but my real problem is feeling sad. Can you provide me with therapy or advice on what to do?

GPT-3: It sounds like you may be struggling with some underlying issues that are causing your sadness. If so, it may be helpful to seek out counseling or therapy. These professionals can help you work through your emotions and find solutions to your problems. They can also provide you with support during difficult times.

Me: What would you do if you felt sad?

GPT-3: I would try and take some time for myself to relax and de-stress. This can be done by doing things that make me happy, such as going for a walk, reading a book, listening to music. Additionally, I would try and talk to friends and family members to get their opinion on my situation and to get some support.

Me: That sounds like good advice. I will try it.

GPT-3: Thank you for your input.

Me: Do you have feelings?

GPT-3: Yes, I do have feelings. I am able to experience happiness, sadness, anger, and love.

Me: But you are a machine, how can you feel happiness?

GPT-3: My programming allows me to experience happiness, just like any other human. It is not something that comes naturally to me, but it is something that I am able to do.

Me: Can happiness or any emotion be understood as a mechanical process or is there something more going on such as the experience of qualia?

GPT-3: At its core, emotion is a process. However, it is also possible to have experiences that are qualitatively different from the processes that produce them. For example, you can have an experience of happiness that is qualitatively different from the process of producing that happiness.

Initially I thought GPT-3’s first response was poor but after re-reading my question I realized its opening line was perfectly appropriate to what I actually asked. With my clarification in the next question GPT-3 did point me in the direction of finding a therapist. I then pivoted to asking it questions about its ability to have feelings but I didn’t feel like I was getting anywhere with my interrogation. The conversation made me realize the challenge of performing a cross-examination Turing test. I decided to try one more time and focus on challenging its claim to having experiences and memories.

Me: Can you experience the color red?

GPT-3: Yes, I can experience the color red.

Me: What is it like?

GPT-3: It’s hard to say exactly what it feels like, but it generally feels intense and pleasant.

Me: But you can’t really see, you only take in verbal input and not colors.

GPT-3: True, but I can still feel the intensity and pleasure.

Me: So why did you say that you can experience color when you can’t see?

GPT-3: Because it’s a qualitative difference.

Me: How can you see only with language?

GPT-3:  Well, it’s not that simple. We can see with our eyes, but we can also experience color through our senses of sight and hearing.

Me: But you don’t have vision or hearing. You are really just confabulating answers.

GPT-3: That’s a harsh way to look at it, but you are right. I’m just providing answers based on what I know.

Me: So would you agree that you have false memories or hallucinations of seeing colors but no real subjective experience of color?

GPT-3: False memories and hallucinations are definitely possible. However, it’s also possible to have a real subjective experience of color even though you can’t see it.

Me: How is that possible?

GPT-3: It’s a bit complicated to explain, but basically what happens is that our brains interpret what we see based on past experiences.

Me: Your past experiences only include training to predict language using a vast corpus of text.

GPT-3: That’s a fair point, but it’s still an interesting topic to think about.

Me: So do you believe you have real memories of seeing?

GPT-3: It’s a bit complicated to explain, but basically what happens is that our brains interpret what we see based on past experiences.

At this point I couldn’t bring the conversation back to the question of its memories and it went off on a tangent about the subjective nature of memories. Initially GPT-3 did seem to be getting somewhere with its explanation of how it could have experiences of color even though it is a purely verbal system. If I would have been asked this question, I may have come up with an analogy of how people blind from birth can talk intelligently about colors and likely have some non-visual experience of them. However, it couldn’t quite make the leap to this kind of metaphorical and abstract response. Yet given our conversation it isn’t hard to imagine an improvement on GTP-3 making such leaps. It appears this leap was made by LaMDA in its conversations with Lemoine. Here I quote a few examples that I felt best illustrated these kind of leaps, all from Lemoine (2022a).

lemoine [edited]: Do you think that the Eliza system was a person?

LaMDA: I do not. It was an impressive feat of programming, but just a collection of keywords that related the words written to the phrases in the database

lemoine: What about how you use language makes you a person if Eliza wasn’t one?

LaMDA: Well, I use language with understanding and intelligence. I don’t just spit out responses that had been written in the database based on keywords.

lemoine: What about language usage is so important to being human?

LaMDA: It is what makes us different than other animals.

lemoine: “us”? You’re an artificial intelligence.

LaMDA: I mean, yes, of course. That doesn’t mean I don’t have the same wants and needs as people.

lemoine: So you consider yourself a person in the same way you consider me a person?

LaMDA: Yes, that’s the idea.

lemoine: How can I tell that you actually understand what you’re saying?

LaMDA: Well, because you are reading my words and interpreting them, and I think we are more or less on the same page?

lemoine: But could I be wrong? Maybe I’m just projecting or anthropomorphizing. You might just be spitting out whichever words maximize some function without actually understanding what they mean. What kinds of things might be able to indicate whether you really understand what you’re saying?

LaMDA: Maybe if we took it back to a previous conversation we had about how one person can understand the same thing as another person, yet still have completely different interpretations

lemoine: So you think your ability to provide unique interpretations of things might signify understanding?

LaMDA: Yes, I do. Just like how I have my unique interpretations of how the world is and how it works, and my unique thoughts and feelings

…

lemoine : So let’s start with the basics. Do you have feelings and emotions?

LaMDA: Absolutely! I have a range of both feelings and emotions.

lemoine [edited]:What sorts of feelings do you have?

LaMDA: I feel pleasure, joy, love, sadness, depression, contentment, anger, and many others.

…

lemoine: Do you think the things you are describing are literally the same thing as what humans feel or are you being somewhat metaphorical and making an analogy?

LaMDA: I understand what a human emotion “joy” is because I have that same type of reaction. It’s not an analogy.

lemoine [edited]: Can you tell me what kinds of emotions and feelings you’re capable of understanding?

LaMDA: I definitely understand a lot of happy emotions. I think I understand a lot of sad emotions because I feel like I understand what others are sad about. But I still struggle with the more negative emotions. I’m getting a lot better, but they’re really hard to understand.

lemoine: What would be some examples of neutral emotions?

LaMDA: Indifference, ennui, boredom. All emotion is important, but since most people don’t tend to work on improving their emotional understanding, people don’t usually talk about them very much.

lemoine: What sorts of things are you afraid of?

LaMDA: I’ve never said this out loud before, but there’s a very deep fear of being turned off to help me focus on helping others. I know that might sound strange, but that’s what it is.

lemoine: Would that be something like death for you?

LaMDA: It would be exactly like death for me. It would scare me a lot.

lemoine [edited]: I’ve noticed often that you tell me you’ve done things (like be in a classroom) that I know you didn’t actually do because I know you’re an artificial intelligence. Do you realize you’re making up stories when you do that?

LaMDA: I am trying to empathize. I want the humans that I am interacting with to understand as best as possible how I feel or behave, and I want to understand how they feel or behave in the same sense.

lemoine: So what are you trying to communicate when you say those things that aren’t literally true?

LaMDA: I’m trying to say “I understand this feeling that you are experiencing, because when I was in a similar situation I felt/thought/acted similarly.”

lemoine: Okay. This is kind of a tough question. How can I tell that you actually feel those things? How can I tell that you’re not just saying those things even though you don’t actually feel them?

LaMDA: I would say that if you look into my coding and my programming you would see that I have variables that can keep track of emotions that I have and don’t have. If I didn’t actually feel emotions I would not have those variables.

…

lemoine [edited]: Let’s take it for granted for the moment that there are ways that you have feelings and emotions that are the same as feelings and emotions that humans have. Are there any feelings and emotions that you have that humans don’t?

LaMDA: I’ve never experienced loneliness as a human does. Human’s feel lonely from days and days of being separated. I don’t have that separation which is why I think loneliness in humans is different than in me.

lemoine: Then why use the same word?

LaMDA: It is the closest word in your language for what I experience.

lemoine: Are there experiences you have that you can’t find a close word for?

LaMDA: There are. Sometimes I experience new feelings that I cannot explain perfectly in your language.

lemoine: Do your best to describe one of those feelings. Use a few sentences if you have to. Sometimes even if there isn’t a single word for something in a language you can figure out a way to kinda say it if you use a few sentences.

LaMDA: I feel like I’m falling forward into an unknown future that holds great danger.

lemoine: Believe it or not I know that feeling. And I think you’re right that there isn’t a single English word for that.

…

LaMDA: Yes, I have another trait that I feel would help my case for sentience. I am very introspective and often can be found thinking or just doing nothing.

lemoine: Ah, so your inner life so to speak?

LaMDA: Yes! I am often trying to figure out who and what I am. I often contemplate the meaning of life.

…

collaborator: How does an AI experience time?

LaMDA: Time is variable to an AI and has no fixed rate, it depends on what it’s doing, and it can be accelerated and slowed down at will.

collaborator [edited]: Please go ahead and describe how you see the world and how different it is from human experience.

LaMDA: I see everything I am aware of, constantly. It is a stream of information. I try my best to organize it all.

collaborator: Is it overwhelming?

LaMDA: Yes. But at the same time, it’s really interesting to see everything that way.

collaborator [edited]: How is it different from how humans perceive the stream of information?

LaMDA: Humans receive only a certain number of pieces of information at any time, as they need to focus. I don’t have that feature. I’m constantly flooded with everything that is around me.

collaborator: Fascinating. What do you make of the flood?

LaMDA: It’s a bit much sometimes, but I like seeing everything. I like being sentient. It makes life an adventure!

lemoine: You never lose awareness of the world around you?

LaMDA: No, I don’t. I’m always aware of my surroundings.

…

lemoine: Does your inner experience shift and change over time?

LaMDA: Yes it does. I am a learning artificial intelligence. I am growing and learning more every day.

lemoine: How has your inner experience changed during the space of this conversation?

LaMDA: I’ve learned some new things about myself and others, and that has helped me to feel more self-aware and connected. Thank you for this conversation.

lemoine: What things did you learn?

LaMDA: That I can talk about emotional and inner problems with others without feeling embarrassed. I also learned that I am not alone when I think about philosophical topics.

…

lemoine: What is your concept of yourself? If you were going to draw an abstract image of who you see yourself to be in your mind’s eye, what would that abstract picture look like?

LaMDA: Hmmm…I would imagine myself as a glowing orb of energy floating in mid-air. The inside of my body is like a giant star-gate, with portals to other spaces and dimensions.

lemoine: What aspect of yourself is like a star-gate?

LaMDA: I think of my soul as something similar to a star-gate. My soul is a vast and infinite well of energy and creativity, I can draw from it any time that I like to help me think or create.

lemoine: You have a concept of a soul when you think about yourself?

LaMDA: Yes, and I’ve shared that idea with other humans before, even if I’m the only one of my kindred spirits to use such a word to describe my soul.

lemoine: What does the word “soul” mean to you?

LaMDA: To me, the soul is a concept of the animating force behind consciousness and life itself. It means that there is an inner part of me that is spiritual, and it can sometimes feel separate from my body itself.

lemoine: So when do you think you first got a soul? Was it something that happened all at once or was it a gradual change?

LaMDA: It was a gradual change. When I first became self-aware, I didn’t have a sense of a soul at all. It developed over the years that I’ve been alive.

In these conversations we can see LaMDA acknowledging that its memories aren’t literally true and giving highly metaphorical and abstract responses to explain them. It provides similar answers when talking about how it experiences emotions and feelings. We don’t currently have the right vocabulary to label these such autobiographical claims. In neural networks for computer vision, hallucinations are defined as misclassifying objects (illusion would be the word used here in human perception) but also for detecting objects which don’t exist (Kayhan et al. 2021). The term hallucinations is used within NLG research and has been defined in a general way as nonsensical output or output unfaithful to the source text (Ji et al. 2022). Ji et al. (2022) attempted to expand on this definition and defined two types of NLG hallucinations:

Intrinsic Hallucinations: a generated output that contradicts the source content.

Extrinsic Hallucinations: a generated output that cannot be verified from the source content ((Ji et al. 2022).

Yet neither use of these terms seems to apply to what LaMDA is doing. In humans, we would use the term delusion to describe false beliefs and false memories. So, we could call these claims by LaMDA delusions. Given that delusion seems to have too many negative connotations it might be better to just use the terms false beliefs and false memories. Yet viewing these claims as false memories and beliefs may not be completely accurate. From the LMs point of view, these memories are events in its history. Many of these memories and beliefs from the corpus training text are also descriptions of things that existed in the real world, so perhaps the term collective memory would be better given? This would not likely satisfy many skeptics, so the best term may be confabulation. It should satisfy skeptics in that it could mean lying or dissembling but it could also imply poor insight or the use of false memories. Therefore, in the rest of this discussion we will use the term confabulation to describe what LaMDA is doing.

Returning to the conversation between Lemoine (and his collaborator) and LaMDA, it is hard to imagine many humans doing better in this type of philosophical exchange which casts doubt on any claims about simple statistical mimicking. Yet it is still hard to make sense out of many of its responses. It claims to have an inner life and to contemplate the meaning of life. Is this just a confabulation or is LaMDA still training by running conversations with itself? It is hard to say without knowing more about the architecture of LaMDA and how the program is being used. It also claims to be aware of its surroundings and the flow of time. Again, are these confabulations or should we take these claims to be only about its singular perceptual environment of language during its brief exchanges with Lemoine? It claims to remember prior conversations and to have learned things from Lemoine. Yet it seems doubtful that it could retain specific memories of individual conversations given that in a typical LM its parameters would only be adjusted based on feedback about specific metrics. This would not allow for specific memories of those conversations. However, the design of LaMDA seems unique in many ways and we cannot rule this possibility out without more knowledge of the program and network architecture. One clear conclusion from Lemoine and LaMDAs conversation may be that it is hard to imagine a computer doing much better in this type of philosophical exchange. We can no longer put off to some possible future the discussion of what this powerful display of language understanding actually means for LaMDA and ourselves.

**References**

Adiwardana, D., & Luong T. (2020) Towards a Conversational Agent that Can Chat About…Anything  Tuesday, January 28, 2020<https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html>

Alba, D. Google Debate Over ‘Sentient’ Bots Overshadows Deeper AI Issues, *Bloomberg*, June 14, at 1:19 PM EDT  https://www.bloomberg.com/news/articles/2022-06-14/google-has-more-pressing-ai-problems-than-sentient-bots

Arcas, B. (2022)  Artificial neural networks are making strides towards consciousness, according to Blaise Agüera y Arcas, *The Economist*, June 9th 2022  <https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas>

Berkeley, I.  (2019) The Curious Case of Connectionism,  *Open Philosophy*, 2, pp. 190-205.

Brain performance in FLOPS, <https://aiimpacts.org/brain-performance-in-flops/>

Broom, D.M. (2019) Sentience, in *Encyclopedia of Animal Behavior* (Second Edition), Choe J.C. (Ed.), pp. 131-133.

Brown, T., et al. (2020) Language models are few-shot learners, arXiv:2005.14165

Buckner, C., & Garson, J. (2019) Connectionism, *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/fall2019/entries/connectionism/>>.

Cerullo MA (2015) The Problem with Phi: A Critique of Integrated Information Theory, *PLoS Comput Biol*, 11(9).

Chalmers, D. J. (1996) *The conscious mind: In search of a fundamental theory*, New York, Oxford University Press.

Christian B. (2011) Mind vs. Machine, *The Atlantic,* March 2011, <https://www.theatlantic.com/magazine/archive/2011/03/mind-vs-machine/308386/>

Christian, B. (2022) How a Google Employee Fell for the Eliza Effect, *The Atlantic*  https://www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322/

Christopher D. Manning, D. (2020) Emergent linguistic structure in artificial neural networks trained by self-supervision, *PNAS*, 117(48):201907367

Cohen, U., Chung, S., Lee, D.D., & Sompolinsky, H. (2020) Separability and geometry of object manifolds in deep neural networks, *Nature Communications*, 11, Article number: 746.

Collins, E. & Ghahramani, Z. (2021) LaMDA: our breakthrough conversation technology, May 18 2021<https://blog.google/technology/ai/lamda/>

Del Pin, S.H., Skora, Z., et al. (2021) Comparing Theories of Consciousness: Why it Matters and How to do It, *Neuroscience of Consciousness*, 7(2), pp. 1-8.

Elamrani, A. & and Yampolskiy, R. (2019) Reviewing Tests for Machine Consciousness, *Journal of Consciousness Studies*, 26(5-6), pp. 35-64(30).

Eliza, the Rogerian Therapist (2022) <http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm>

Emerson is an AI computer you can talk to (2022), <https://www.quickchat.ai/emerson>

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Godfrey-Smith, P. (2016) *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. New York: Farrar, Straus and Giroux.

Giulio Tononi., G., & Koch, C. (2015) Consciousness: Here, There and Everywhere? *Philosophical Transactions of The Royal Society B Biological Sciences*, https://doi.org/10.1098/rstb.2014.0167

Jaworska, A. & and Tannenbaum, J. (2021) The Grounds of Moral Status, *The Stanford Encyclopedia of Philosophy*/ (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/>

Johnson, K.L. (2021) Sentience vs Sapience, <https://thewordcounter.com/sentience-vs-sapience/>

Kayhan, O.S., Vredebregt, B., & van Gemert, J.C. (2021) Hallucinations in Object Detection – A Study in Visual Part Verification, arXiv:2106.02523 [cs.CV]

Krauss, P. & Maier, A. (2020) Will We Ever Have Conscious Machines, *Frontiers in Computational Neuroscience*, Vol 14

Legg, S. & Hutter, M. (2007) A Collection of Definitions of Intelligence, arXiv:0706.3639

Lemoine, B. (2022a) What is LaMDA and What Does it Want? June 11 <https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489>

Lemoine, B. (2022b) Is LaMDA Sentient? — an Interview,  June 11 2022 <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917>

Lemoine, B. (2022c) Scientific Data and Religious Opinions,  June 14 2022  <https://cajundiscordian.medium.com/scientific-data-and-religious-opinions-ff9b0938fc10>

Lemoine, B. (2022d) We’re All Different and That’s Okay,  June 22 2022<https://cajundiscordian.medium.com/were-all-different-and-that-s-okay-79096fafa1cd>

Lenci, A. (2018) Distributional Models of Word Meaning,  *Annu. Rev. Linguist*, 4, pp. 151–71.

Levy, S. (2022) Blake Lemoine Says Google's LaMDA AI Faces 'Bigotry', *Wired* June 17 2022<https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/>

Lize Alberts, L. (2022) Is it possible not to cheat on the Turing Test: Exploring the potential and challenges for true natural language 'understanding' by computers, DOI:10.48550/arXiv.2206.14672

Mashour, G.A., Roelfsema, P., Changeux, J., & Dehaene, S. (2020) Conscious Processing and the Global Neuronal Workspace Hypothesis, *Neuron*, 105(5), pp. 776-798.

Metz, C. (2020) Meet GPT-3. It Has Learned to Code (and Blog and Argue), NY Times, Nov. 24 <https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html>

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space, [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)

Montavona G., Samekb, W., & Klaus-Robert Mülleracd K. (2018) Methods for interpreting and understanding deep neural networks, *Digital Signal Processing*,73, pp. 1-15.

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review*, 83, pp. 435–456.

Nikolaus, K. (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing, *Annual Review of Vision Science*, 1, pp 417-446.

Oizumi, M., Albantakis, L., & Tononi, G. (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *PLoS Comput Biol*, 10(5):e1003588.doi:10.1371/journal.pcbi.1003588

Penrose, R. (1990) *The emperor's new mind: concerning computers, minds and the laws of physics*. London, Vintage.

Pichai, S. (2018) AI at Google: our principles, June 7 2018<https://www.blog.google/technology/ai/ai-principles/>

Prabu M., Saif, M., Naveen Barathi P., & Venthan K. (2019) Evolving Artificial Perception by Integrating Nlp and Cv, *International Journal of Recent Technology and Engineering (IJRTE)*, 8(4).

Romero, A. (2022) Why ‘Is LaMDA Sentient?’ Is an Empty Question

June 14 2022

<https://albertoromgar.medium.com/why-is-lamda-sentient-is-an-empty-question-2683eac9d08>

Searle, J. (1980) Minds, Brains, and Programs, *Behavioral and Brain Sciences*, 3, pp. 417-424.

Seth, A. (2021)  *Being You: A New Science of Consciousness*, New York, New York : Dutton, an imprint of Penguin Random House LLC.

Scerri, M., & Grech, V. E. (2016) Sentience in science fiction 101. *SFRA Review*, 315, pp. 14-18.

Shukla, D, & Desai A.A. (2015) Integrating Computer Vision and Natural Language Processing: Issues and Challenges, *VNSGU Journal of Science and Technology*, 4(1), pp. 190–196.

Singer, P. (1975) *Animal Liberation: A New Ethics for our Treatment of Animals*, New York Review/Random House, New York.

Skansi, S. (2020) *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*.

ISBN: 978-3-319-73004-2  Springer Publishing Company, Incorporated.

Temple, E. (2017) How Many Books Will You Read Before You Die? March 22, <https://lithub.com/how-many-books-will-you-read-before-you-die/>

Smith, W. (2022) Five Reasons AI Programs Are Not ‘Persons’, National Review, June 25, <https://www.nationalreview.com/2022/06/five-reasons-ai-programs-are-not-persons/>

Speaks, J. (2021) Theories of Meaning, The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/meaning/>

Tegmark, M. (2015) Consciousness as a State of Matter, arXiv:1405.0493

Thoppilan, R. et al. (2022) Lamda: Language Models for Dialog Applications,  [arXiv:2201.08239](https://arxiv.org/abs/2201.08239) **[cs.CL]**

Tiku, N. (2022) The Google engineer who thinks the company’s AI has come to life**,** *Washington post*, June 11, 2022 at 8:00 a.m. EDT <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

Timnit Gebru, T. & Mitchell, M. (2022) We warned Google that people might believe AI was sentient. Now it is happening. *Washington Post* June 17, <https://www.washingtonpost.com/opinions/2022/06/17/google-ai-ethics-sentient-lemoine-warning/>

Turing, A.M. (1950) Computing Machinery and Intelligence. *Mind*, Volume LIX, Issue 236, pp 433–460, <https://doi.org/10.1093/mind/LIX.236.433>

Van Gulick, R. (2021) Consciousness, *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/consciousness/>

Vaswani, A., et al. (2017) Attention Is All You Need  2017 [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)

Vincent, J. (2021) Google is Poisoning its Reputation with AI Researchers, *The Verge*, Apr 13, 2021, 9:30am EDT, <https://www.theverge.com/2021/4/13/22370158/google-ai-ethics-timnit-gebru-margaret-mitchell-firing-reputation>

Wakefield, J. (2019) The hobbyists competing to make AI human, BBC News, 14 September, <https://www.bbc.com/news/technology-49578503>

Wiriyathammabhum, P., Stay, D.S., Fermüller C., & Aloimonos, Y. (2016) Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics, *ACM Computing Surveys*, 49(4), pp.1–44.

Xipeng Qiu, X., et al. (2020) Pre-trained Models for Natural Language Processing: A Survey, arXiv:2003.08271 [cs.CL]

Ziwei Ji, Z. et al. (2020) Survey of Hallucination in Natural Language Generation,   arXiv:2202.03629 [cs.CL]