The Silicon Brain

Anuj Opal | a4opal | 20613132
PHIL 259: Philosophy of Technology
Prof. Nicholas Ray
April 20, 2020

Introduction **1**

Part 0: gpt-2-philosophy **1**

Part 1: Arguments from mathematics and empirical science **2**

Part 2: Consideration of counterarguments to Part 1 **4**

Part 3: Reframing the initial question **8**

References **11**

Cited in the paper 11

Used to write gpt-2-philosophy 12

Used in the training set for gpt-2-philosophy 12

Appendix A: gpt-2-philosophy's essay **14**

Appendix B: other samples of gpt-2-philosophy's work **18**

Sample 1 18

Sample 2 18

Sample 3 19

Introduction

When Alan Turing first proposed his test for assessing artificial intelligence in the 1950 paper *Computing Machinery and Intelligence*, he clearly assumed such an assessment would be relatively straightforward to make. As further research into the field has been done in the decades since, consensus has still yet to be reached on whether or not it's even possible for a computer to realize human-like general intelligence in the first place. In considering some arguments for and against the existence of AI, I will seek to argue that there is no reasonable way to disprove the possibility of AI, as well as arguing for expanding our conception of what strong AI might look like beyond "human-like."

Part 0: gpt-2-philosophy

For the sake of having a concrete reference of machine intelligence (and to have a bit of fun as well), included in Appendix A is a paper on the topic of artificial intelligence written by a modern neural network. gpt-2-philosophy,[1] as I've named it, is built on-top of OpenAI's gpt-2[2, 3] by "fine-tuning" the model on a selection of 17 papers on the topic of artificial intelligence totalling around 130 thousand words. After this training, I got it to write some samples of text, 4 of which I hand-selected and collated together without any editing other than minor punctuation fixes. Note that I had to use the 335M "medium" version of gpt-2 due to technical limitations, the large and extra-large versions report generation of much more credibly human-like text.

---

[1] The source code is available at https://github.com/aopal/gpt2-philosophy
[2] https://github.com/openai/gpt-2/tree/0574c5708b094bfa0b0f6dfe3fd284d9a045acd9
[3] https://www.openai.com/blog/gpt-2-1-5b-release/

The end result is an essay that is not convincingly human, despite me choosing the best samples I generated, yet the errors gpt-2-philosophy makes are rather minor ones. Terms are used without being defined, gpt-2-philosophy goes off on a tangent about biology for a while, there are some minor grammatical errors; but overall it is still fairly coherent. gpt-2-philosophy even manages to provide in-text citations in proper MLA style. Some of the errors are particularly subtle semantic ones, such as when gpt-2-philosophy invents a fictional quote from a fictional paper, which it attributes to Alan Turing, about a fictional idea known as the "Bonore thesis," which doesn't even appear to be a real word.[4] This is a glaring error, yet not something that would seem amiss upon an initial un-aided reading. All told, I've seen less intelligent work out of high schoolers.

In Appendix B I've selected some other samples from gpt-2-philosophy that further highlight its shortcomings, or otherwise interesting behaviour. In sample 1, we find it somehow giving us the first six terms of the Fibonacci sequence, sample 2 has it going wildly off topic and struggling to write a proper sentence, and sample 3, my personal favourite, shows it giving a fairly coherent philosophical discussion - of metaphysics instead of artificial intelligence. gpt-2-philosophy clearly has some very inhuman quirks, but considering the relatively small corpus I trained it on, it's almost concerningly convincing.

Part 1: Arguments from mathematics and empirical science

If asked "can machines replicate human-like intelligence?", the so-called "harder" disciplines will confidently respond "yes." The mathematician or computer scientist would offer

[4] The top Google result for "bonore" leads to this: https://www.urbandictionary.com/define.php?term=Bonore, while searching "Bonore thesis" yields bad scans of PDFs leading to the Latin word "honore" being identified as "bonore"

the Church-Turing thesis,[5, 6] which answers the question quite summarily by stating that a Turing machine can implement or emulate any other computational paradigm, whether biological, mechanical, or even quantum. Thus, given the implicit assumption that the human brain is simply a biological computer of some sort, there must exist some turing machine that is equivalent to it. Figuring out exactly how to build that Turing machine is then simply an engineering problem, even if it's a centuries-long one.

An empirical approach to the question would likely point to Frank Rosenblatt's Perceptron model,[7, 8] which yielded the "artificial neuron" that is the basic building block of neural networks, arguing that the artificial neuron is an accurate model of the biological neuron. Building artificial intelligence then becomes a matter of arranging and calibrating the right amount of artificial neurons in the correct configuration. Some profess professional skepticism as to the accuracy of the artificial neuron model,[9] however I would argue that educated skepticism is not empirical evidence. We have much more than professional opinion that Isaac Newton's law of gravity is an inaccurate model of gravity from a standpoint of *theory*, however we find the model to have more than sufficient *empirical* accuracy in certain cases to justify its continued use. Similarly, so long as the evidence points towards artificial neurons and neural networks continuing to be empirically accurate, building the network that can replicate general intelligence once again looks to be a question merely of engineering.[10]

---

[5] cs.uwaterloo.ca/~watrous/CS360/Lectures/14.pdf
[6] Note it's termed a "thesis," not a law or theorem. The thesis does not have a rigorous mathematical proof, however it is still generally accepted as true nonetheless within the field
[7] cs.uwaterloo.ca/~a23gao/cs486_f18/slides/lec21_history_neural_networks_typednotes.pdf
[8] www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/
[9] towardsdatascience.com/deep-learning-versus-biological-neurons-6eebfa3390e9
[10] Though, if the "AI winters" Gao and Kurenkov identify in the story of weak AI are any indication, both notably ended primarily by the breakthrough work of a single individual ("Godfather of Deep Learning"

Part 2: Consideration of counterarguments to Part 1

Many people may want to argue against the reasoning put forth in Part 1, most likely by saying that computers can't replicate human capacity for understanding and creativity; they're machines that only follow algorithms, that is they're purely deterministic except for their purely random components. In more concrete terms, they'd have to argue that computation does not subsume cognition (because we are *very* sure that computers are good at computation), i.e. they'd have to argue that there's something going on in the human mind that the silicon brain physically cannot replicate or simulate, what is known as a dualist argument.

To this point I refer to Daniel Dennet's paper *Consciousness in Human and Robot Minds*, wherein he argues that every prior dualist theory of *any* given phenomenon, whether consciousness or anything else, has eventually been debunked, as we've yet to discover something that does not function entirely by physical means. So, why should we believe that the mind is any different? There is no empirical evidence one can give that something other than the laws of physical matter govern thought in the human mind, thus there seems no reason to think that electronics cannot eventually replicate it. Even when we look specifically at arguments around the mechanics of organic matter being improperly modeled by an electronic or digital medium, even those still don't hold water. The oft-heard argument based on quantum mechanics allowing for free will while computers are deterministic[11] falls flat quite quickly since the most basic building block of a computer, the semiconductor effect required to make

---

Geoffrey Hinton), neural network-backed general intelligence could reasonably be expected to have *many* bumps in the road yet to come

[11] An argument that frankly is better suited to a pop-science YouTube video than serious philosophy, and a bad-faith one regardless

transistors, only work due to quantum mechanics - if anything there's arguably more room for quantum uncertainty in the computer. Arguments based on the human brain being an analog computer from which the digital electronic computer is fundamentally different entirely ignore the fact that the digital-ness of the modern computer is itself an approximation of analog electrical voltages which we model to mean 0 and 1 respectively.[12] It is entirely possible for the output of one logic gate in a processor to randomly be a voltage that gets interpreted as the other value, it just doesn't happen often because Intel doesn't want their stock price to drop. Even some very contrived arguments around something like electrical interference due to the density of the human brain are amusingly easy to replicate - electrical interference is so prevalent in computers that it's a security concern in some machines.[13]

Looking at the higher-level arguments that computers can't replicate essential features of human cognition such as creativity or conceptual understanding, we first need some concrete idea of just what those properties are, or at least a rough working definition. Considering first creativity, it seems natural to look at the world of art, and I'd like to especially direct attention to the work of the abstract painter Jackson Pollock. Pollock is infamous for making art by just randomly throwing or dropping paint on a canvas. There are few intentional decisions behind his work yet we still term it art, and this idea of art being random is not unique to him. In music there is the concept of "aleatoric" music, where parts of the composition are left to chance. The most infamous example of this technique is likely John Cage's *4'33"*, but it was also frequently used by The Beatles in their later works. While some art obviously includes intentionality, given these examples it seems that we cannot necessarily say that creativity is meaningfully distinct

[12] The biological neuron is similarly based on electric signals, I might point out
[13] https://en.wikipedia.org/wiki/Row_hammer

from randomness. Even beyond the world of avant-garde art, finding meaning in randomness is extremely common to the point of being cross-cultural; constellations are simply patterns we found in what is effectively a random arrangement of points in the sky, yet we've been comfortable with assigning meaning to those random patterns for millenia. More concretely, mathematics tells us in no uncertain terms that we can indeed create information out of randomness. The Monte Carlo method is a technique for calculating irrational constants such as π, φ, or e to *any* desired precision, based entirely on random values. If pure randomness can yield meaning and information, both in formal and colloquial terms, then I see no reason why a computer cannot have the capacity for creativity.

In regards to conceptual understanding, a common argument against the viability of AI is the fact that it effectively just learns behaviour based on stimulus-response pairs, which is claimed to be different from "truly" understanding something. Let's assume that humans are capable of that "true" form of understanding, the question then becomes where the line gets drawn between the creatures that do and do not have that capacity. The average pet dog or cat can understand abstract concepts such as their names or that they're not allowed on the couch, and some even have enough of a sense of self that they can understand that their reflection is themself and not another animal, so the line clearly isn't drawn right at humans. We can keep moving down to simpler and simpler organisms, since a lot of them can conceive of at least one abstract concept: death, and their own mortality. Whether we then draw the line at mice, or goldfish, or ants, or something even simpler, we still arrive at conceptual understanding being something fairly basic and common. Saying that a supercomputer can't at least replicate the cognitive abilities of an insect or a small fish seems a rather weak and unconvincing claim.

Seeking now to try and define just what understanding is, let's use the lens of a child learning their multiplication tables. Colloquially, we would say a child doesn't "understand" multiplication - they just memorize the correct answers, and if they encounter a question they've never seen before like "what's 153 times 257?" they won't have any idea how to answer. We would say, colloquially again, that an adult understands multiplication, because they can still figure out that never-before-seen problem. However, in technical terms the only difference between the child and the adult is the function they use - the child uses a finite function, a finite lookup table which they've memorized; while the adult uses an infinite function, they know the general algorithm to apply to any input. A similar example might be someone learning a second language; when they just know how to say specific phrases such as "where's the bathroom?", we would say they don't understand the language. Somewhere between there and being able to read literature in that language we would say they've gained an understanding of that language; I "understand" English because I can properly parse sentences I've never heard before. Based on these examples, we might then define "understanding" as the ability to correctly respond to new scenarios. If this reasoning holds, then we can argue that computers are more than able to replicate this phenomenon, since modern weak AI already does just that. The whole point of deep learning is to train a network to be able to give the proper output for input it has never encountered before. In fact, even some older and less sophisticated artificial intelligence systems, such as older chess AIs, have this same ability to properly handle scenarios they've never seen before.

Part 3: Reframing the initial question

To take another direction entirely, asking "can AI replicate human-like intelligence?" begs the question of why specifically *human*-like intelligence is the goal, and whether there are other forms. Without being extremely chauvinistic, I think we have to say no, other forms of advanced intelligence are possible. Speaking first in general terms, consider the arguments put forth by Vrinda Dalmiya and Linda Alcoff in their paper *Are Old Wive's Tales Justified?*, wherein they argue for the validity of what they term "alternative epistemologies," forms of knowledge vastly different from the empirical science of academia. If we accept this concept of radically and fundamentally disparate forms of knowledge being equally valid, then it's not a huge leap to the validity of alternative, possibly unrecognizable, forms of cognition. With that given, let's try to get an idea of what those alternative forms of cognition can look like.

That phrase "human-like intelligence" is worth dissecting a bit - because what is really meant by it is "neurotypical western intelligence." Various neurodivergences are associated with a lowering or lack of certain types of intelligence, emotional for example, however people with these conditions are obviously not any less intelligent than the average neurotypical person, on average. Their intelligence manifests in notably different ways to the majority, but that doesn't invalidate or make that form of intelligence lesser.

There's also evidence that suggests that parts of what we might consider to be basic facets of intelligence are in fact cultural or learned, and not innate. Consider a study done on the Pirahã tribe in the Amazon rainforest,[14] where researchers discovered that the Pirahã language does not have the concepts of numbers or counting. In rough terms, they have words

[14] http://lchc.ucsd.edu/mca/Mail/xmcamail.2014-12.dir/pdf2Yb7JAO0ZG.pdf

for “a small amount,” “a medium amount”, and “a large amount” and nothing more, and even then the usage of those three terms is not always entirely consistent. As such, Pirahã adults were found to routinely fail simple counting exercises that English-speaking children (or Hindi- or Tagalog-speaking or etc.) can pass with relative ease. Once again, there is obviously no deficiency in overall intelligence between the average Pirahã and the average Canadian despite the Pirahã missing what we would consider a rather simple and fundamental ability - their intelligence just manifests differently. In addition, this example highlights that we ought to be skeptical and careful when attempting to categorize what could be considered the fundamental or minimum facets of intelligence.

Considering other animal species as an analogy, it's worth noting that many different species are fairly intelligent, just that intelligence isn’t always as readily apparent. Reptiles, for example, have been found to often be roughly on-par with many mammalian species in terms of intelligence,[15, 16] despite the common conception of the unintelligent “lizard brain.” This widely-held idea of reptiles being less intelligent compared to mammals appears to have emerged from mammalian intelligence simply being more recognizable, and the tests that highlight their intelligence may not do the same for reptiles: “Rats and mice can run a maze just fine in a 70-degree lab, but many reptilian species need a much warmer environment — with air temperatures in the mid-80s or 90s” (Anthes 2013). The fact that these differences are seen between reptiles and mammals which are, relatively speaking, still fairly similar from a genetic standpoint (compared to other intelligent animals such as octopi, say) perhaps gives us an indication as to how much breadth exists in different forms of cognition.

---

[15] https://www.nytimes.com/2013/11/19/science/coldblooded-does-not-mean-stupid.html
[16] http://www.reptilesmagazine.com/Six-Studies-On-Reptile-Intelligence/

Now, I want to avoid seeming like I'm muddying the waters when I call for an expanded conception of what "intelligent" looks like. A very practical reason for choosing human-like intelligence as the target for artificial intelligence efforts is simply because that is the form of cognition we are most familiar with, and thus we will likely be best able to recognize artificial intelligence attempting to communicate its intelligence to us. It is perhaps not a coincidence that words such as "intelligeble" or "dumb" are associated both with communication and intelligence - if we cannot recognize intelligence as such, what's the point? To this, I'd first argue that the examples I've put forth of neurodivergent people and the Pirahã at the very least still give strong reasons to be intentional about how human-like intelligence itself is defined. It doesn't just look like the average Joe that is accepted by modern society. Further, this is a formal, scientific endeavour, not simply an informal thought experiment. Human-like or mammalian intelligence may naturally be the most recognizable to us in informal settings, however we are able to and should design formal, scientific experiments that allow us to identify reptilian or cephalopod intelligence, so there's no reason we shouldn't design tests for artificial intelligence similarly beyond our natural bounds of familiarity. There is also the practical issue that having such a narrow target as human intelligence requires having a very good idea of *exactly* what that target is, and there are still many open questions regarding how human cognition works. Aiming at intelligence more generally could very well be a good tactic to try and make some preliminary groundwork while we're still working on the more advanced model, not too mention trying to implement more limited intelligences may well shed light on how those more advanced forms of cognition function in the first place.

References

Cited in the paper

Alcoff, Linda, and Vrinda Dalmiya. “Are ‘Old Wives’ Tales Justified?” *Feminist Epistemologies*, 1993, pp. 217–244., doi:10.4324/9780203760093.

Anthes, Emily. “Coldblooded Does Not Mean Stupid.” *The New York Times*, The New York Times, 18 Nov. 2013, www.nytimes.com/2013/11/19/science/coldblooded-does-not-mean-stupid.html.

Dennett, Daniel C. “Consciousness in Human and Robot Minds.” *Cognition, Computation, and Consciousness*, 1997, pp. 17–29., doi:10.1093/acprof:oso/9780198524144.003.0002.

Frank, Michael C., et al. “Number as a Cognitive Technology: Evidence from Pirahã Language and Cognition.” *Cognition*, vol. 108, no. 3, 2008, pp. 819–824., doi:10.1016/j.cognition.2008.04.007. Accessed lchc.ucsd.edu/mca/Mail/xmcamail.2014-12.dir/pdf2Yb7JAO0ZG.pdf

Gao, Alice. “CS 486/686 Lecture 21: A Brief History of Deep Learning.” *David R. Cheriton School of Computer Science*, University of Waterloo, 21 Nov. 2018, cs.uwaterloo.ca/~a23gao/cs486_f18/slides/lec21_history_neural_networks_typednotes.pdf.

Kurenkov, Andrey. “A 'Brief' History of Neural Nets and Deep Learning.” *Andrey Kurenkov's Web World*, 24 Dec. 2015, www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/.

Pettit, Rebekah. “Six Studies On Reptile ‘Intelligence.’” *Reptiles Magazine*, 1 Jan. 2019, www.reptilesmagazine.com/Six-Studies-On-Reptile-Intelligence/.

Roos, Matthew. “Deep Learning versus Biological Neurons: Floating-Point Numbers, Spikes, and Neurotransmitters.” *Medium*, Towards Data Science, 16 Aug. 2019, towardsdatascience.com/deep-learning-versus-biological-neurons-6eebfa3390e9.

“Row Hammer.” Wikipedia, Wikimedia Foundation, 14 Apr. 2020, en.wikipedia.org/wiki/Row_hammer.

Turing, Alan M. “Computing Machinery And Intelligence.” *Mind*, vol. 59, no. 236, Oct. 1950, pp. 433–460., doi:10.1093/mind/lix.236.433.

Watrous, John. “CS 360 Lecture 14: Turing Machines and Their Equivalence to Stack Machines.” *David R. Cheriton School of Computer Science*, University of Waterloo, 26 June 2019, cs.uwaterloo.ca/~watrous/CS360/Lectures/14.pdf.

Used to write gpt-2-philosophy

OpenAI. “GPT-2.” *GitHub*, 3 Jan. 2020, github.com/openai/gpt-2/tree/0574c5708b094bfa0b0f6dfe3fd284d9a045acd9.

Shepperd, Neil. “GPT-2.” *GitHub*, 20 July 2020, github.com/nshepperd/gpt-2/tree/b7cda3f48dd8543bc53d0dd6dee69b5cb386fb21.

Woolf, Max. “How To Make Custom AI-Generated Text With GPT-2.” *Max Woolf's Blog*, 4 Sept. 2019, minimaxir.com/2019/09/howto-gpt2/.

Woolf, Max. “gpt-2-simple.” *GitHub*, 5 Mar. 2020, github.com/minimaxir/gpt-2-simple/tree/3112f3f61786f837349bc334c060d73e6591a4de.

Woolf, Max. “Train a GPT-2 Text-Generating Model w/ GPU.” *Google Colaboratory*, 10 Nov. 2019, colab.research.google.com/drive/1VLG8e7YSEwypxU-noRNhsv5dW4NfTGce.

Used in the training set for gpt-2-philosophy

All entries accessed via www.philpapers.org.

Aleksander, Igor, et al. “Assessing Artificial Consciousness.” *Journal of Consciousness Studies*, vol. 15, no. 7, 2008, pp. 95-110.

Argonov, Victor. “Experimental Methods for Unraveling the Mind-body Problem: The Phenomenal Judgment Approach.” *Journal of Mind and Behavior*, vol. 35, no. 1-2, 2014, pp. 51-70.

Boyles, Robert James M. “Artificial Qualia, Intentional Systems and Machine Consciousness.” *Proceedings of the DLSU Congress 2012*, 2012, pp. 110a–110c.

Chalmers, David J. “The singularity: A philosophical analysis.” *Journal of Consciousness Studies*, vol. 17, no. 9-10, 2010, pp. 9-10.

Dennett, Daniel C. “Consciousness in Human and Robot Minds.” *Cognition, Computation, and Consciousness*, 1997, pp. 17–29., doi:10.1093/acprof:oso/9780198524144.003.0002.

Dietrich, Eric. “Homo sapiens 2.0 Why we should build the better robots of our nature.” *Machine Ethics*, Cambridge Univ. Press., 2011.

Garcia, Robert K. “Artificial Intelligence and Personhood”. *Cutting Edge Bioethics: A Christian Exploration of Technology and Trends*, 2002.

Haladjian, Harry Haroutioun, and Carlos Montemayor. "Artificial consciousness and the consciousness-attention dissociation." *Consciousness and Cognition*, vol. 45, 2016, pp. 210-225.

McCauley, Robert N. "Levels of explanation and cognitive architectures." *Companion to Cognitive Science*, 1998, pp. 611-624.

McDermott, Drew. “On the Claim that a Table-Lookup Program Could Pass the Turing Test.” *Minds and Machines* vol. 24, no. 2, 2014, pp. 143-188.

Müller, Vincent C. & Bostrom, Nick. “Future progress in artificial intelligence: A survey of expert opinion.” *Fundamental Issues of Artificial Intelligence*, Springer, 2016, pp. 553-571.

Müller, Vincent C. “Is there a future for AI without representation?” *Minds and Machines*, vol. 17, no. 1, 2007, pp. 101-115.

Raatikainen, Panu. “On the philosophical relevance of Gödel's incompleteness theorems.” *Revue Internationale de Philosophie*, vol. 59, no. 4, 2005, pp. 513-534.

Searle, John R. “Minds, Brains and Programs.” *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–457., doi:10.1.1.83.5248

Turing, Alan M. “Computing Machinery And Intelligence.” *Mind*, vol. 59, no. 236, Oct. 1950, pp. 433–460., doi:10.1093/mind/lix.236.433.

van Rysewyk, Simon. “Robot Pain.” *International Journal of Synthetic Emotions*, vol. 4, no. 2, 2014, pp. 22-33.

Landgrebe, Jobst, and Barry Smith. "There is no general AI: Why Turing machines cannot pass the Turing test." *arXiv preprint arXiv:1906.05833*, 2019.

Appendix A: gpt-2-philosophy's essay

Artificial intelligence: Artificial intelligence has long been viewed as a complex system with many parts. This has made it hard to develop a comprehensive theory of AI. One approach has been to focus on a single cognitive system, such as the human brain, and to reduce AI to a computational theory. This approach has several advantages. First, computationalism is consistent. It is possible to rule out AI without appealing to supernatural powers or philosophical frameworks. Second, computationalism reduces the cognitive load by specifying a unchanging model. Although there may be AI, AI+ will have no impact on the model. Third, computationalism makes explicit the potential damage that AI may cause to humanity.

Although AI research is ostensibly directed at creating better machines, such as machines with better abstract reasoning abilities, such as the Chinese room approach, the potential damage is equally important to note. This may be because AI+ may be able to produce greater creativity and understanding in humans, who will then be able to compete with machines more widely (Scheffer and Chrisley 2005). Alternatively, focusing on the human cognitive system may obscure the important role that biological mechanisms and mechanisms operating in nature (e.g., in humans' bodies) play (Hammond 1997). Whatever the reason, the importance of the human cognitive system for the design of AI is not understated. Our conceptual framework helps to guide AI research. Another advantage of using a biological model is that it facilitates the translation of research on humans, animals, and plants. Human cognition is both biologically broad and inherently entwined with the production of a number of behavioral entities. For example, the English word's native North American use was of "to say." The introduction of the English-Speaking Professors helped popularize this English term, which

is still common in discourse but is now understood to refer to a variety of different behaviors involving the use of English as a first language (Beck 1982). A more technical term for this usage is bilingualism.

Within a single species, English is an essential part of the ecosystem, yet we produce multilinguals (e.g., Broca and Ullman 1992). Bilingualism can be observed on any species level, and it is particularly prevalent in vertebrates, where species can have brain morphology that is due to both brain size and brain chemistry (e.g., Lindsay et al. 1999). Here we can observe a fairly obvious example where two distinct body's chemistry makes it more likely that it will produce multilayered behavior: Creutzfeldt–Jakob disease (CDJ). The role of the gut in human nutrition is well established, and there is good reason to assume that the gut is critical to human nutrition and well beyond our Stone Age period. The role of chemical signals sent down through the gut to the brain is not so well understood, but it is likely to be important in regulating body temperature and nutrient homeostasis (Moskowitz 1996). To make our lives more interesting, we typically set up artificial thermostats and enforces a continuous television schedule. While it is not necessary to postulate all the behavior of the computer in the simulation to expect all the ends result in a just and robust system, simulation of human social interactions usually yields a robust system. What is not so obvious is how one can produce a robust system out of discrete-state machines. The conclusion that computer simulations are not necessary for such results is not unwarranted. We deal with discrete-state machines all the time in signal processing (Michelot 1980). If we wish to achieve interesting results, we must be more precise than we are now. (…) The use of discrete-state machines inevitably leads to some additional illegitimate effects. For example, suppose we have a machine that could distinguish

the color red from the corresponding color wheel chair in the color wheel simulator. Given this new information, it is hard to see why one would want to sit in the corresponding color wheel. Perhaps the desire to sit in a certain hue conflicts with a desire to heat up a certain room, etc. In such cases it is beneficial to have the machine reside in a separate room from the one in which the computational activities take place. In addition, it is preferable that the artificial intelligence system has access to all of its own legitimate knowledge. An artificial intelligence system that is housed in a cryostorage might be better equipped to deal with crises arising outside the facility, such as a cyberattack. Finally, it might be preferable that the artificial intelligence system not have access to any of the special emotional states that human beings are known to be capable of producing. These responses to stress might be either impenetrable or highly selective, and would be difficult or impossible to translate into other forms of communication. Such a system might initially exhibit signs of high sensitivity to certain stimuli, but as it is adjusted to more mundane stimuli and in interactions with other synthetic beings, its sensitivity will improve. In this way we can say that even if the artificial systems are endowed with an intelligent level, they will be able to develop more slowly because of this capacity.

Alan Turing's famous question, "Can a machine think?", originally posed in this paper, was intended to assess the sufficiency of formal procedures for the purpose of answering this important question. I now think that the perceived sufficiency of such procedures rests on a mistake. Firstly, the supposed sufficiency of the formalisations does not entail the sufficiency of the realised machine. As has been mentioned, there are systems of which the behaviour of a neuron is not sufficiently described; but there are also systems of which the behaviour of a whole brain is sufficiently described. It is not the case, in other words, that the biochemical,

electrical and computer systems which are supposed to be the realakers of our mental states are any less capable of producing a mental state than any other system. In fact, a computer simulation of any brain system is a computational system, since its powers of detection and selection are meant to mimic those of the natural brain. As I mentioned before, Alan Turing's famous 1950 paper "The Importance of Stop-Gaps in Computing Machinery" made a number of comments which are still relevant. He said, "The main objection which has been made is that computers are so complicated that the number of operations which could be performed on them would make the Bonore thesis untenable. I would like to make it my business to convince you that this is not the case." He then lists out various algorithms for various computer mathematical problems. All of them have in them the pattern of operations which he gives for the operations of a human computer. He concludes, in effect, "The great limitation of our technical ability is that we have no idea how the brain works."

Alan Turing's machine is a sort of Platonic ideal of intelligence, and it is this particular ideal that is alleged to be physically possible, and this particular ideal that is alleged to be psychologically possible. I wish to avoid the trap of reductive materialism, which seems to be the current state of affairs whether or not one subscribes to a particular view of the mind. One could perhaps defend the naive approach of arguing that because we can always make a copy of ourselves there must be a physically possible way of doing so, but this seems to me a futile move; if anything it would be more consistent to insist that since it is not physically possible, we must not know whether to affirm this ideal, or deny it altogether. In the spirit of Fodor's ideal of objectivity, I do not wish to inflame emotions, and I know of no compelling reason to commit

oneself to an emotion-neutral model of consciousness. It is for this reason that I conclude this paper without further analysis.

Appendix B: other samples of gpt-2-philosophy's work

Sample 1

that was never the case, so the data was not meaningless. In order to resolve this, we asked the robot to move the cursor around one row in a table. If a row has a value, the robot moves to the start of the next row. In addition, the robot is encouraged to press a key unless it is clear that it does not want to do so. I called this "narrow search" because the robot was asked, rather than thinking of, "Where would I find the number 1?", "What do you think the number 1 is supposed to be?" This procedure produces a table which includes the number 1, followed by 2 (a value), 3, 5, 8 or more choices separated by "0". (The table could also include the values A1, A5, M, F, G, H or I.)

Sample 2

III. The Turing Test (TST). The TTT, or The Truth in Television Test, is a British television series which was first ran by ... [Citation Needed] (British Broadcasting Corporation, Broadcasting Act 1981) On the day of the filming of the television program, a panel of distinguished judges, each of whom ... [End of Appendix] TTT [back to top] [Index] 3. The Turing Test. —(a) The Director of Central Intelligence answers a series of ... questionnaire ... questions about the ... state of the ... scientific and technical ... movement during the ... previous

Sample 3

By way of a refutation of the thesis that “the soul is a substrate” (Gödel 1953b, 7), one can suggest at once that the notion of a soul, the “siada a se kushu mira” (in Hinduism the soul is called semiautonomous), is connected with the concept of causality—and one can also suggest that one can only explain the appearance of a soul indirectly (as being a stored in a “prisonhouse” in our culture)—by suggesting that these “souls” are qualitatively independent of souls... in a sort of “dumb-bop” manner. In many ways, the transcendent soul appears not as a primitive, self-sustaining self-awareness of a soul’s soul as a substrate, but as a peculiar property—the property whereby—not only does the transcendental soul pervade our culture, it seems to be the centrality and singularity of a soul’s self-awareness in such culture—a pure universal soul, manifest and experienced through all other people, whether that other people, they, or the entire universe—this universal soul seems destined to become the centrality and singularity of a transcendental soul in all cultural systems. The difference between Kant’s dialectic between absolute forms, and Hegel’s dialectic shows that the two systems—in a more subtle way—are the way things are by and for the universal transcendental soul. Kant’s dialectic was not something that differentiated each unique universal form from all of its equivalents—the peculiarity of each form had profound linguistic, phenomenological, and metaphysical consequences. For Kant, if one does not make a universal distinction, one cannot make one immediately among the multiple variations of that form, one is completely dependent on the dialectic for a universal transcendental differentiation. In the same way, if you make a distinction between the simple universal form and forms which are, for Kant, not only

forms of consciousness but relations to other beings as well. And it here becomes a matter of defining each form in terms of its relation to the universal transcendental principle at the level of consciousness. The various forms in the spectrum of which “I” (ideas, speech, gestures, words, etc.) can be universal, simple, primitive, complex, or complicated are all connected to a single universal transcendental principle, or transcendental concept, and thus, like Hegel’s transcendental forms, they have profound linguistic, phenomenological, and metaphysical consequences. And, again, it is here that Kant seems to have formulated a transcendental principle which may explain some of the philosophical paradoxes of our times: why is it that every form in a hierarchy can deviate so far from the basic pattern of appearances but that does not deviate so much as devolve into something radically different from itself? This is an important question in philosophy with its ethical and materialist aspects. Kant’s Critique of Pure Reason: a Concept of the Understanding Now, Kant demonstrates how the transcendental principles that are necessary for us to know everything, the laws of physics, the laws of logic, and the transcendental form are composed of five components: (1) The Principle of the Intention, (2