Producing Human-Level Dialogue with Transformer-Driven AI Philosophers

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper examines a verbal discussion of three artificial intelligence agents, each described by a generative language model. Each agent's language model is fine-tuned separately from a baseline DistilGPT2 model to give each a unique personality corresponding to the philosophers Plato, Friedrich Nietzche, and Lao Tzu. Each has an internal memory used to generate new sentences, which correspond to the psychological concepts of "thoughts", and they are periodically allowed to communicate, by letting one of their thoughts be broadcast into the others' memories. Qualitative analysis of sampled results are promising, showing the models achieve an approximation of their respective personalities and the communication has some impact upon their respective sequential generative processes. There is still significant room for improvement, including a better method for quantitative analysis and assessment.

1 Introduction

One of the primary drives toward the development of artificial intelligence (AI) as a subfield of computer science was the promise of building intelligent machines, i.e., programs that could think with the flexibility, creativity, and internal drive that characterizes human intelligence. While recent dramatic leaps have been made in the subfield of machine and learning, less focus has been given toward such a realization of artificial general intelligence, as noted by early pioneers of the pre-AI winters era.[1] One open-ended question that has led to this decline in focus is the ambiguity of the problem. Bolstrom has identified three pathways to superintelligence, a similar goal as artificial general intelligence: (1) speed superintelligence, (2) collective superintelligence), and (3) quality superintelligence.[1] The prospect of collective superintelligence is examined here, whereupon a large quantity of human-level or sub-human-level artificial intelligences may communicate and *collectively* attain a baseline intelligence. The task then becomes how to design one intelligent machine and how to design an effective communication interface between them. This paper will examine how the recent incredible success in generative language modeling with deep learning methods may have leveraged for this purpose.

The first task at hand is to define a so-called "thinking machine". The psychological thinking process is, in the general current scientific paradigm, unmeasurable.[2] However, certain functions or transforms of thoughts are measurable, namely via spoken or written language and brain measurements. Here written language as a function of thought will be examined; that is:

$$\text{language} = f(\text{thinking}) \tag{1}$$

In the spirit of scientific analysis[3], thinking may be broken down into constituent "thoughts", which may be treated artificially as discrete. In this case, it may be postulated that a correspondence exists such that:

$$\text{sentence} = f(\text{thought}) \tag{2}$$

Then a human-based thinking machine may be thought of as one such that it can time-step from one thought to the next in a manner similar to humans. This is almost exactly one-to-one with the requirements established for artificial cognition by the Turing test.[4] What is gained by such an approximation is an abundance of data, as there are at least 130 million books available and a large, uncounted corpus of human-generated language on the internet.[5] As such, deep learning methods hold great promise in being able to accurately construct such a sentence time-stepper. This promise is due to two reasons: (1) Neural networks are universal function approximators, and (2) the quality of this approximation is data-driven, i.e., if more data (of the same distribution as what is trying to be approximated) is given to the network, it generally will improve the approximation. However, neural networks represent a kingdom of various domains, phyla, orders and species. Care must be taken to select a proper network architecture for the task at hand, as they are not universally applicable in all cases.

The first distinction in the search for a proper architecture is the need for a generative model. Given past inputs, the model must generate an output. This is a regression task, but in its capacity to do this repeatedly it is called *autoregressive*, also called a generative model. This aligns it with the mechanics of a standard time-stepper, such as the Runge-Kutta fourth order method, which constructs a new data point given previous data points and dynamics of the system. Second, the model must be a "language model", which is to say it approximates human language. As such, it will need to take in linguistic data, usually discretized into written words instead of spoken sound data. Prior to being input into any such model, such a sequential stream of words can be quantified into one-hot vocabulary tokens, and then transformed into word embeddings to latently encode the semantic relationships between words. These embeddings are their own separate task often performed using neural networks, such as the word2vec system[6], yet may be taken as given for the task at hand, as they do not generally require retraining. This constructs as input for the system a matrix, with each word being a vector, stacked together in the order of the sentence.

Here an important architectural decision is made. Recurrent neural networks (RNNs) were once a common choice, though they have fallen out of favor due to lower performance. Convolutional neural networks (CNNs) hold potential promise here, in applying generative methods such as WaveNet or Pervasive Attention, but neither have yet been implemented in an accessible way for word-sequence-to-word-sequence tasks.[7][8] A third method, the transformer, was invented in 2014 for the specific purpose of natural language processing (NLP), borrowing from the insights of attention from encoder-decoder RNN architectures.[9] They have rapidly become the hammer of choice for a wide variety of methods, including many of the most important recent breakthroughs in deep learning, such as GPT3 and AlphaFold2.[10][11]

The transformer is a highly engineered architecture, though there are fundamental subunits that can be analyzed to understand its function: (1) Self-attention, and (2) A feedforward layer. The feedforward layer is a standard, fully connected layer in a neural network. The novelty of the transformer comes from the self-attention mechanism. The intuitive purpose of this process is to transform the independent word embeddings into a more holistical "sentence embedding". Let us first consider the input, the word embeddings. This uses some pretrained embedding, designed to capture the latent semantics of a standalone word. This is remarkably successful, with the famous example being the preservation of the meaning when the linear operation "King $-$ Man $+$ Woman $=$ Queen" is performed. However, the interest of a transformer is to analyze a sequence of words, such as a sentence. In the construction of a sentence, each word is related to other ways in new, latent ways not captured by the independent words. A classic example is trying to decipher to which other words the word "it" is related to. In this manner, the self-attention block tries to re-encode each of the word vectors to match their meaning within that particular sentence, or sequence of input text. This is done by describing each word as a weighted sum of all of the word vectors. These weights are calculated by a series of projections using trainable "query", "key", and "value" matrices. In some cases, this might place most of the weight (or "attention") on the word itself, if the original word embedding does indeed capture the desired semantics. But for words like "it", and many other pronouns, conjunctions, articles, etc., most of the weight/attention will be given to the words they refer to. As such, the self-attention block, the core building block of the transformer, may be seen as the logical extension of word embeddings to sentence embeddings, though still outputting a sequences of vectors. These are called attention "heads". For this reason, they are seen to be an excellent choice of architecture for the original purpose of constructing a sentence time-stepper. The feedforward network then acts upon these new vectors. Multiple self-attention/feedforward blocks are often stacked on top of each other, allowing non-linear effects to be integrated. One further innovation of the transformer architecture is the use of multiple attention heads for each self-attention block, allowing multiple different attention schemes to be concatenated and mixed together via one last (trainable) linear transformation. Equating each attention head with a human, which presumably is the metaphorical origin of the terminology, this is akin to having multiple people, each with their own internal dictionaries and internal linguistic schemes, come together and collectively determine the best sentence embedding. Now that the general methodology of the transformer has been explained, it is important, though somewhat tangential, to note that the original word embeddings are also summed with a positional encoding vecor, which is usually some form of frequency space information, such that the original information is minimally corrupted yet by examining frequency information of the embedding the position can still be inferred from the network. There are, of course, numerous other details, such as the novel use of layer-normalization instead of batch normalization, and standard details such as using residual layers and softmax after self-attention layers. Yet these critical elements of the architecture distinguish transformers as being in their own class of model. [9][12]

Many variants of the transformer architecture rapidly proliferated, often to accommodate specific NLP tasks. Generative pre-trained transformer 2 (GPT2) gained wide-spread attention for its use in generative language modeling tasks, surpassing previous successes by a wide margin, based upon subjective assessment, and even instigating concerns over the ethical implications of its ability to generate realistic content. The key distinguishing feature of GPT2 is the use of only decoding blocks. The original transformer block used two self-attention blocks in tandem; theoretically, little is lost by only using one, with one exception. The decoder block is only allowed to "attend" (assign weights to) words prior to it in the sequence; it cannot peak into the future of the sentence.[13] This is important for the task of text generation, though less so for tasks related to language understanding. For this reason, models such as Bidirectional Encoder Representations from Transformers (BERT) use only encoders.[14] GPT2 took advantage of the fact that sentence-generation models, and time-series forecasting models in general, are unsupervised in the sense that they do not require labeled data. Unlabeled sequences of data are used and the last element of the sequence becomes the target data point of the model. As such, OpenAI first crawled Reddit for 8 million web pages and over 40GB of internet text, still a vastly small subset of the total internet text corpus.[15] The original GPT2 model has 1.5 billion parameters, though the smallest version of the model, DistilGPT2, is now only 82 million parameters.[15][16] This reduction was performed using knowledge distillation, in which a smaller neural network is "taught" by a larger neural network.[17]

One motivation behind this work was the advent of GPT3, the follow-up work by OpenAI from GPT2, which has come remarkably close to passing the initial phases of a Turing test.[18] GPT3 has 175 billion parameters, eliciting concern over the growing inaccessibility of such models due to immense computational resources and environmental concerns due to proportional electricity demand.[10] Such progress in sentence time-stepping technology promotes questions of if it is time to turn to the question of the role of such a technology as the "engine" in a larger artificial general intelligence system.

2 Methods

Using transformers in an applied application encounters an immediate problem. While the number of hours used to train the original model is unknown, comparable models such as BERT had computational costs due to training in the tens of thousands of dollars. As such, the paradigm has shifted in the transformer-based applied NLP community to transfer learning or fine-tuning pre-trained models.[19] Transfer learning is used when a slightly different task is needed than the pre-trained model, while fine-tuning is used when the same task is used, but it is desired to train the model on new content. For example, GPT2 is trained from web content crawled from Reddit, so its generated content has a similar texture and focus of content. Fine-tuning GPT2 has been used to, for example, make a text generator that generates Marxism tweets , showing the flexibility of the model and the specific powers of fine-tuning.[20] The next choice is which model to fine tune. The smallest available model, DistilGPT2 was chosen, due to its size. Since the purpose of this work is to create a society of transformer-driven agents, each of which will need its own model, as small of a footprint as possible is preferred to allow for future expansion of this work into larger numbers of agents. Additionally, DistilGPT2's model is approximately 315 GBs, letting it fit within the RAM of a Raspberry Pi Zero, important for embedded applications.

Access to and interfacing with DistilGPT2 is available through the Huggingface API *transformers* library, which provides an internal model storage usable with both PyTorch and TensorFlow. Training can be performed on raw text files. For classic books whose copyright is expired or non-extant, such text files can be found at Gutenberg.org. With these tools in place, the exact setup of the "society" model of the experiment can be discussed.

The basic setup is to conceive of an artificial mind as a point in some sort of mental space, where each point is represented by a sentence. The transformer model transforms word embeddings into sentence embeddings, and uses internal feedforward neural networks to step to the next sentence, by generating one word at a time. This core represents the "engine" of each AI agent. Moving forward, "thought" will be used interchangeably with "sentence" for the AI system. However, a collective intelligence requires communication channels, much like how each human has significant regions of its physiology and neuroanatomy designed specifically for communication, such as Wernicke's area, Broca's area, and the human vocal system. Let the following approximation/postulate be made: language is an approximation of telepathy, that is, the direct broadcast of thoughts from one mind to another. Let each artificial intelligence have a memory of each thought/sentence it has ever output before, yet be only able to hold one thought in its mind at once. Then, much like the standard distributed computing paradigm, such a society may undergo "computation/communication" cycles, where during the "thinking" phase, equivalent to "computation", each AI agent uses its specific fine-tuned sentence stepper to create a new thought, taking its entire memory as input. At the "communication" phase, *one* of the AI agents thinks a new thought via its sentence time-stepper, and this is *broadcast* to every other AI, becoming its new state. This might be considered *memetic recombination*, in contrast to genetic recombination, where time-propagating minds interchange ideas (memes), in metaphor with the exchange of genes in organism reproduction time steps. The AI that generates the new thought is the "speaker", and every other AI is the "listener". This is an approximation of communication in a society. It is noted that significant improvements can be made to this model, but as a baseline experiment this represents a first approximation. After this communication phase, the think/communicate cycle repeats, for as many times as desired. This setup is shown in Figure 1.

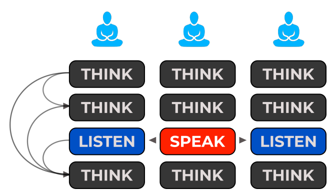

Figure 1: A schema for the think/communicate cycle. Two "thinks" are done, then one "communication". An additional think is placed at the end to show the entire cycle can repeat.

Note that the AI agent is always taking its entire past history as input. The DistilGPT2 model can handle 1024 tokens, which are approximately proportional to words, after which the memory acts as a queue and kicks out the oldest "memories", the first sentences it trained on.

With this setup detailed, the specific trained models will be discussed. First, a goal for the society is established. While a metric for the quality of an artificial general intelligence is difficult, being the problem that originated the Turing test, perhaps one interesting metric is the capacity of a *group* of artificial intelligences to accomplish a goal that is deemed "AI-complete".[1] As such, it would be beneficial to have a group of AI with transformer time-stepper thinking engines fine-tuned on a specific domain of knowledge, with each trained for diversity within this domain. While a scientific or technological domain would be preferable, philosophy is used as a stand-in domain for the following reasons: (1) Equations present a challenge to handle, (2) Philosopher's often have a very unique style or "philosophy" that can be easily identified to a reader acquainted with their works, as a subjective metric of success, (3) While philosophy is significantly less tangible than science or technology, it is still feasible to imagine a group of sufficiently sophisticated artificial intelligences developing a "new" philosophic theory with merits over extant human-originated ones, as evaluated by human experts in the field. Compare with how a group of AI physicists might discover an equation for dark energy. Such a feat will surely not be done here, as this is an introductory attempt at this method, yet such an opportunity is an important element to keep in mind. Three agents were used, to add the jump in complexity from a two agent system. Three corpuses of text were used, one for each model associated with each agent: (1) The Republic, by Plato, (2) Tao Te Ching, by Lao

Tzu, and (3) Thus Spoke Zarathustra, by Friedrich Nietzsche. Each is meant to be representative of the author in general. The Republic concerns justice, Plato's theory of forms, and the role of the philosopher in society.[21] Tao Te Ching is the foundational text of Taoism, and emphasizes the virtues of naturalness and non-action.[22]. Thus Spoke Zarathustra stresses the classical virtues of nobility, pride, and victory over the medieval virtues of humility, meekness, and altruism.[23] These brief descriptions are mentioned to allowed the reader to subjectively assess the ultimate quality of the model.

Each of these texts was split into an 80% training set and an 20% validation set using cross-validation. Severe dataset size constraints were imposed, in part to see how the model could train on such constrained datasets of various sizes. This allows insight into how rapidly the DistilGPT2 model can adapt and how little information is needed to approximate a given personality. The translations used for The Republic has 118,383 words, Thus Spoke Zarathustra has 89,909 words, and Tao Te Ching has 10,664 words. A linearly decaying learning rate scheduler with warm-up was used with the AdamW optimizer, with initial learning rate 5×10^{-5} and (β_1, β_2) parameters of $(0.9, 0.999)$.[24][25] The number of epochs used was experimented with.

Generative language models, such as chatbots, face an ultimate problem of being difficult to quantitatively assess. This is, in part, due to their goal being to simulate human cognition, which is not well-defined. This difficulty is seen in the Turing test, by definition requiring human assessment. With this said, the models themselves establish loss via a perplexity score, an information-theoretic construct that measures how well a probability model predicts a sample. A lower score indicates the model is better at predicting the sample.[26] The perplexity scores show only how well the agent can reproduce the text, not how well the generated output will appear to human observers. Furthermore, the central idea of the society is to promote the memetic recombination of thoughts, of which it is not presently known by the author how best to analyze. As such, one method here will be to sample one *random* society cycle, and evaluate the results subjectively. In the future, it is hoped to apply classic time series data analysis methods, such as the analysis of the periodic spatial modes using dynamic mode decomposition, where here the "space" is the word space constructed from the word embeddings. Such a method would allow for both a much more rigorous treatment, and to analyze hundreds or thousands of such cycles with varying lengths of think/communicate phases. Even with these difficulties, reporting perplexities scores and subjective content analysis seems standard for comparable generative language model analyses.[27] [1]

3 Results

The average perplexity scores over five cross-validation trials were:

Name	Perplexity (0 Epochs)	Perplexity (15 Epochs)	Perplexity (200 Epochs)
The Republic	66	46	1024
Tao Te Ching	70	50	855
Thus Spoke Zarathustra	106	64	1020

Each of the models achieved their best average perplexity scores around 15 epochs. However, it is noted that this is only a slight improvement from the zero epoch perplexity score, representing the models with no fine tuning. Conversely, each model scores extremely high perplexity score with 200 epochs. This is in contrast to subjective inspection, which found the models to be best in the 200 epoch regime. As such, even though lower perplexities were found in the low epoch regime, 200 epochs were used for the following qualitative analysis. The author has not found a clear explanation for this discrepancy, though one possible explanation is as follows. Small datasets were used in each case, and the 1/5 validation set was taken contiguously, thus always removing 1/5th of the logical flow of the work, which is not guaranteed to be well represented in the remaining 4/5ths. However, this does not explain why the 200 epoch models did not score better than the zero epoch model on the sole account of similar word choice. For example, the word "Tao" is used frequently in Tao Te Ching, yet would not be expected to be used at all in the model with no fine tuning. It is also possible that there are underlying mechanics of the perplexity score not yet understood by the author accounting

[1] Code used is available at: `https://github.com/benfrancis314/amath_563`.

for this discrepancy. In any case, the subjective analysis shows a clearer personality corresponding with what is expected based off of the text with the additional training.

The society schedule shown in Figure 1 was then conducted. Each model began with the same seed sentence as a prompt, "What is the meaning of life?". The results are shown in Figure 2.

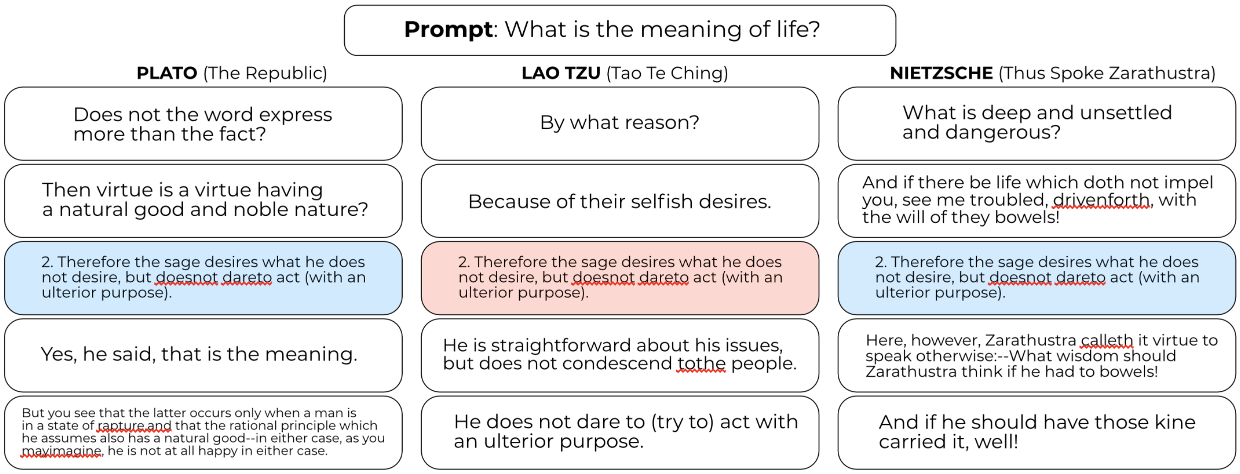

Figure 2: A random sample of a trialogue between the three agents with the prompt "What is the meaning of life?". One communication phase and two thinking phases are shown. A longer version is given in the supplementary material.

The initial prompt is met by responses from each model, which take as input only the prompt. The models are named after the author of the text used for training. Before analyzing, the reader is once again cautioned toward the following subjective analysis; it represents a sort of Rorschach test that is with complete certainty shaped by the subjective opinions of the author. However, as described above, subjective analysis is an important element of evaluating generative language models, and so it will be done with as much care as spatial constraints allow. This is not a definitive scoring of the results, but rather an example of how one *can* parse through and try to see how the model performed.

Plato responds with a remark on the difference between names and facts, which seems related to his central "Theory of Forms", which relates to the difference between perception and ideas.[28] Lao Tzu responds with a question, fitting with the general koan-like nature of Taosim (from which Zen Bhuddism in part derives).[29] Nietzsche responds with a question relating the meaning of life to what may easily be taken as an allusion to the dangerous depths of the unconscious, a concept the influence of his work helped create.[30] At this point, each philosopher AI agent has shown a certain degree of the personalities they were trained to replicate. Each then responds only to themselves; Plato's new input is both the prompt *and* his first thought. He remarks something about virtue, nobleness, nature, and good, which is roughly fitting with The Republic's thematic concepts. However, it does seem disconnected, not building upon the prompt. Lao Tzu seems to answer his first thought as if it was a question for himself to answer, instead of a clarification question to the prompt. It does not quite make sense, but the concept of selfish desires as the negative is fitting with what the Tao Te Ching warns against, with the book being generally a juxtaposition of the wise sage and the unenlightened acts of the general populace. Nietzsche responds with fervor, in a slightly incoherent sentence, yet which upon second reading does mention life, the prompt, and promotes life that internally impels oneself. This is conceivably an answer to the prompt, though again only through the lens of interpretation. The second clause is generally nonsensical, perhaps confusing the exact meaning of the word "bowel". Perhaps not. At the third time step, Lao Tzu is randomly chosen to be the speaker. He speaks something circular about desiring and inaction. An artifact of the training set is also seen, where the Tao Te Ching teachings are written with numbers in front. Additionally, there are minor spelling issues, missing spaces where there should be. This is broadcast to Plato and Nietzsche; they do not sentence time-step this round, and next round their input will be a concatenation of [the prompt, thought 1, thought 2, Lao Tzu's speech]. Then each of their next responses might be considered their "responses", though mechanically it only operates that a foreign sentence has been inserted into their memory. Plato responds directly like one would in a dialogue, saying "Yes, he said, this is the meaning". This is perhaps an artifact of the fact that The Republic is itself a dialogue, and so Plato is predisposed to generating text like a dialogue. It also shows the long memory of the transformer model, being able to reach back to the prompt for the word "meaning". Nietzsche responds by essentially disagreeing, which is a sign of a well-adapted personality, as Lao Tzu's statement is basically antithetical to the morality presented in Thus Spoke Zarathustra. This is perhaps the most important time block of the trialogue, as it shows that, in this random sample, the models were able to interact with each other via the communication mechanism, to some degree; the recombination had effects lasting for at least one time step, in a manner somewhat similar to how human discourse might following one philosopher presenting an idea to others. Another time step is shown, to show that while the immediate time step from the exchange was promising, the next time step is less so. Plato intriguingly makes a "but" statement, but it is hard to decipher what it is referring to, though there is reference to a "latter" statement, of which Lao Tzu did have two clauses in his speech. As a whole, it is still not quite coherent. Lao Tzu essentially repeats the second clause of his speech, showing a tendency Lao Tzu's model has to get stuck in loops. Nietzsche makes a somewhat non-sensical statement with regard to context, though the texture or personality is fitting in with that of the "biblical oratory and playfulness" style of Thus Spoke Zarathustra.[23]

An extended version of this trialogue is available in the supplementary material. This analysis shows how the model was successful in some regards, and showed promise for the communicative effects of the recombination step. Furthermore, it showed a distinct personality fit to each philosopher agent, seeming to match the training text. In other elements, there was a lack of cohesion within the time frame of five time steps.

4 Conclusion

In conclusion, this work examined an adapted version of a Turing test in which AI agents talk with each other, and used a small version of GPT2 to generate the language for such a discussion, fine-tuned on three separate models to give them unique "identities". In the subjective examination of a small sample of this discussion, there was both promise that the model setup can lead to interesting results, and a clear need for improvement. Chief among the next steps for improving this model is a better form of analysis, to guide changes beyond slow and biased human subjective analysis. The first candidate for this is to adapt classical time series analysis methods, such as DMD, time-delay embedding, etc., for the output in word-embedding form, such that it will be a series of latently meaningful vectors. An additional option would be the use of Amazon's Mechanical Turk to accumulate enough subjective analysis for the use of statistical methods.[31] Four further possibilities for improvement are: (1) Reach into the transformer architecture and make tweaks. There is limited flexibility due to the immense constraints on pretraining time, but transfer learning could be done by adjusting the edge layers. (2) Train the sentence time-stepper on additional content to give it a more fleshed out personality; it might be beneficial or interesting to train it upon the life works of an actual living person, such as the voluminous extant work of Theodore Roosevelt. (3) Surround the agent with additional architecture. For example, construct a reinforcement learning environment that it could interact with, and use the sentence time-stepping as only one constituent component of its mind, perhaps generating multiple candidate sentences at once and comparing their results with other elements of its cognitive architecture. This highlights the question of what the role is of deep learning in an artificial mind; which components should be learned, and which components should be engineered? (4) Make a more sophisticated society model. For one thing, many more agents could be used, limited by the 300 MB footprint. Then a random subset of them could be speakers, and an exclusive random subset of the non-speakers could be the listeners for each speaker. In tandem with the reinforcement learning idea, an environment could be constructed to let the society as a whole "do something", perhaps considering it as a whole to be one collective agent, similar to how the brain is composed of intelligent neurons, and in fitting with the overarching goal of collective intelligence or superintelligence. When GPT3 is available to the general public, this may represent an opportunity to swap out the sentence time-stepper used, though the same size constraints will need to be considered.

References

[1] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014.

[2] J. Friedenberg and G. Silverman. *Cognitive Science: An Introduction to the Study of Mind*. SAGE Publications, 2015.

[3] F. Wilczek. *A Beautiful Question: Finding Nature's Deep Design*. Penguin Press, 2015.

[4] A. M. TURING. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, 10 1950.

[5] Leonid Taycher. Books of the world, stand up and be counted! all 129,864,880 of you.

[6] Jay Alammar. The illustrated word2vec.

[7] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[8] Maha Elbayad, Laurent Besacier, and Jakob Verbeek. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. *CoRR*, abs/1808.03867, 2018.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[11] Dale Markowitz. Alphafold 2 explained: A semi-deep dive.

[12] HuggingFace. The illustrated transformer.

[13] HuggingFace. The illustrated gpt-2.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] Wikipedia contributors. Gpt-2 — Wikipedia, the free encyclopedia, 2021. [Online; accessed 9-June-2021].

[16] HuggingFace. Distilgpt2.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[18] Kevin Lackler. Giving gpt3 a turing test.

[19] Sebastian Ruder. The state of transfer learning in nlp.

[20] Mohamad Ali Nasser. Step-by-step guide on how to train gpt-2 on books using google colab.

[21] Wikipedia contributors. Republic (plato) — Wikipedia, the free encyclopedia, 2021. [Online; accessed 9-June-2021].

[22] Wikipedia contributors. Tao te ching — Wikipedia, the free encyclopedia, 2021. [Online; accessed 9-June-2021].

[23] F.W. Nietzsche, C. Martin, K.M. Higgins, and R.C. Solomon. *Thus Spoke Zarathustra*. Barnes & Noble Classics. Barnes & Noble Classics, 2005.

[24] Sylvain Gugger. Adamw and super-convergence is now the fastest way to train neural nets.

[25] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[26] Wikipedia contributors. Perplexity — Wikipedia, the free encyclopedia, 2021. [Online; accessed 10-June-2021].

[27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[28] Wikipedia contributors. Theory of forms — Wikipedia, the free encyclopedia, 2021. [Online; accessed 9-June-2021].

[29] Wikipedia contributors. Zen — Wikipedia, the free encyclopedia, 2021. [Online; accessed 9-June-2021].

[30] C. Jung. *Psychological Types*. Routledge Classics. Taylor & Francis, 2016.

[31] Amazon. Amazon mechanical turk.