Limits of AI: Using Kant's Transcendental Idealism to Find the Limitations of Transformer Models for Policy Making, Computer Science, and Epistemology

Bachelor Philosophy of a Specific Discipline Thesis
Words: 11000

Student: Dyami van Kooten Pássaro (532415)
Supervisor: Prof. Dr. Y Hui
Advisor: Prof. Dr. W. van Bunge
Main Study: Systems Engineering, Policy Analysis, and Management (TU DELFT)
Date: 15/07/2024

Acknowledgements and Foreword

I'd like to thank my supervisor, Prof. Dr. Yuk Hui, for his guidance on the content of the thesis and helping me take my first steps toward rigorous Philosophy. I'd also like to thank my tutor, Ruby Knipscheer, for her guidance on the form of the thesis. Lastly, I'd like to extend a special thank you to Ward Kint, my highschool Philosophy teacher, for his enthusiasm which got me into this field in the first place.

I had a wonderful time exploring the intersection of my various interests in the Epistemology of Artificial Intelligence, and hope you may experience as much joy reading this thesis as I had writing it.

Table of Contents

1.1: AI's rise to the public consciousness

In late November 2022, Sam Altman's OpenAI released a product that would catapult Artificial Intelligence (AI) from a vague term used in technology circles to the forefront of public consciousness[1]. Within a few months of the release of GPT-3, it was clear that these Large Language Models (LLMs) would change the way in which institutions like schools or workplaces would conduct their activities moving forward.

While it had been clear for a number of years that Machine Learning (ML), a subsection of AI, was increasingly changing the world with its applications, the human-like conduct of GPT-3's chatbot ignited speculation on whether we were close to achieving something close to Artificial General Intelligence (AGI), i.e. a model that could replicate human intelligence and possibly even surpass it. This was compounded by the nature of the chatbot-interface, masking the technical aspects behind the algorithm and leaving only an interface that could seem to, in a certain sense, possess human-like intelligence and reasoning abilities.

1.2: Concerns about AI and calls for regulation

This prompted a strong reaction from citizens and regulators alike, with concerns being raised about the effects such a technology could have on society at large, including existential threats such as those seen in movies like *The Terminator*[2]. This commotion gave way to increased regulatory scrutiny and even calls from industry leaders on a halt to AI training[3].
The reasoning behind this proposed halt comes from a concern regarding the rapid improvement of these models and the supposed irreversible negative consequences they could bring before regulators have time to catch-up on the technology and consider these consequences properly.

Furthermore, according to Kadtke and Wells[4], the rate of technological change is not merely speeding up, but in fact accelerating, making it even more difficult for regulators to assess how such a technology will develop, let alone what it's consequences will be. This exponential development would in itself not be an issue if ample time was available for regulators to assess before negative consequences were to occur. Unfortunately, this is unlikely to be the case, due to the so-called "Collingridge Dilemma" or "control dilemma" posited by David Collingridge in his 1980 book *The Social Control of Technology*[5].

---

[1] Marr, Bernard. "A Short History of ChatGPT: How We Got to Where We Are Today." *Forbes*, February 20, 2024. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/
[2] Watercutter, Angela. "Imagine if Joe Biden's AI Executive Order Were Inspired by 'The Terminator.'" *WIRED*, November 3, 2023. https://www.wired.com/story/imagine-if-joe-bidens-ai-executive-order-were-inspired-by-the-terminator/
[3] Vallance, Chris. "Elon Musk Among Experts Urging a Halt to AI Training," March 30, 2023. https://www.bbc.com/news/technology-65110030
[4] Kadtke, James B., and Linton Wells. *Policy challenges of accelerating technological change: Security policy and strategy implications of parallel scientific revolutions*. Center for Technology and National Security Policy, National Defense University, 2014.
[5] Collingridge, David. "The Social Control of Technology," 1980. https://repository.library.georgetown.edu/handle/10822/792071

1.3: The Collingridge Dilemma and its implications for regulators in AI policy-making

The Collingridge Dilemma states that as a certain technology becomes more entrenched into society, it becomes increasingly more difficult to regulate or control it in such a way as to negate its unwanted effects. This can occur due to a wide-ranging set of factors, from the subtle changes technologies bring to our moral frameworks[6] to more straight-forward factors of companies and consumers simply starting to rely on a given product or service. As these effects are often hard to anticipate, and will cause a dependency on the technology despite its negative effects, regulators will often be too late in realizing the damage that is being done until the technology is already too deeply rooted for them to exert significant influence on its development or use.

This implies that regulators will tend to be 'behind the curve' not only in not understanding the technical specifications and the uses of new technologies, but also their longer term, undesirable effects. It is for this reason that academics such as Van de Poel feel that simply anticipating technological trends is not sufficient, and deeper philosophical work is needed[7], which he does through his 'Social Experiments'[8], while Kudina and Verbeek suggest their own alternative in the form of techno-moral scenarios.

Still, these solutions work best when a technology has a relatively stable identity against which these solutions can be applied to in order to alleviate the dilemma, as a detailed risk-analysis becomes difficult with quickly changing use-cases stemming from an exponential developmental trajectory. As such, a thorough analysis of these AI models is needed that tries to figure out this underlying identity. This will be done by looking not at how incremental increases in computing power may help the models develop, but rather at the structure of the models itself, in order to figure out how far we can reasonably assume current architectural designs may take us, regardless of compute.

1.4: Moving towards an identity

Thus we have arrived at the main point of interest of this thesis: With the introduction of a new technology or use-cases such as those of LLMs, regulators will tend to act too late to prevent negative consequences due to the Collingridge Dilemma. While some literature exists on how to tackle this problem, it is mostly pointed towards technologies that already have a fairly stable identity or development trajectory, where scenario-analysis and experimentation can help. With LLMs such as GPT-3 already having an incredibly big impact early in their development-cycle, and with this development not only being exponential, but likely accelerating in pace, we need a more fundamental approach based on inherent limitations. This would allow us to not only set more realistic expectations of what such a model may be able to achieve over time, but also inform the creation of new models to move us more into a direction that allows AI to be more capable.

---

[6] Kudina, Olya, and Peter-Paul Verbeek. "Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy." *Science, Technology, & Human Values* 44, no. 2 (2019): 291–314. https://www.jstor.org/stable/26637439

[7] Van De Poel, Ibo. "An Ethical Framework for Evaluating Experimental Technology." *Science and Engineering Ethics* 22, no. 3 (November 14, 2015): 667–86. https://doi.org/10.1007/s11948-015-9724-3

[8] Van De Poel, Ibo. "Why New Technologies Should Be Conceived as Social Experiments." *Ethics, Policy & Environment* 16, no. 3 (October 1, 2013): 352–55. https://doi.org/10.1080/21550085.2013.844575

1.5: Kant's Transcendental Idealism for achieving the identity

Comparisons towards this purpose of structural analysis to system limitations can be made with Immanuel Kant's Critique of Pure Reason[9], which introduced his Transcendental Idealism, and looking at the limits of human cognition (i.e. epistemology) through the structure of its experience itself. This approach to epistemology is often viewed as joining the traditions of rationalism and empiricism into a coherent system, whereas they were largely in opposition before. This may be relevant towards AI in that an analogous divide can historically be seen between the work on Symbolic and Connectionist Systems, roughly mirroring the former philosophical traditions according to Goel[10]. Where Symbolic system represent the world through relations of logical objects, akin to rationalism's focus on a priori cognition, the Connectionist approach is more concerned with creating structure out of data, in line with the empiricist approach. As such, Kant is the logical starting point for a transcendental analysis of AI-systems in line with the goals of this thesis. Comparing his system of human cognition with their AI-counterpart, we may hope to gain insights into what the limitations of these models are, and how they may (not) be able to become truly intelligent through generating knowledge.

Given the constraints on this thesis in both time and expertise, an all-encompassing analysis of AI is not achievable, not least of which because it is an umbrella term that encompasses a lot of different approaches and model types. Rather, given that LLMs have recently been the cause of AIs surgent popularity, we will look at LLMs, and more specifically the Transformer Architecture, through the lens of Kant's Transcendental Idealism, in order to find its limits. This choice of architecture is not arbitrary, as specifically the Transformer Architecture has been the bedrock of the new wave of generative AI breakthroughs in the LLM-space. Furthermore, it should be noted the thesis rests on an assumption that mapping such systems over human intelligence is optimal, and leaves out the analysis based on different structures, as it falls outside of the scope of what is achievable in one analysis. Thus, our research question is as follows:

"What are the structural limitations of Transformer models' ability in generating knowledge according to Kantian Transcendental Idealism?"

1.6: Thesis Structure

The thesis will be structured as follows: We will first look at Kant's Transcendental Framework so we can understand the elements that are needed for human-like intelligence and knowledge generation in section 2. We will then analyze the Transformer Architecture to see how that lends itself to world-creation, i.e. try to establish what the subjective apperception of such an algorithm may look like, in section 3. Following this, we will contrast these systems against each other to spot any differences or missing functions in Transformer-based systems, including touching on their inherent limitations, in section 4. Finally, we will elaborate on what implications these conclusions have on Policy Making, Computer Science, and possibly even Epistemology, in section 5.

---

[9] Kant, Immanuel. *Critique of Pure Reason*. *Cambridge University Press eBooks*, 1998. https://doi.org/10.1017/cbo9780511804649

[10] Goel, Ashok K. "Looking Back, Looking Ahead: Symbolic Versus Connectionist AI." In *The AI Magazine/AI Magazine* 42, no. 4 (December 1, 2021): 83–85. https://doi.org/10.1609/aaai.12026

2: Kant's Transcendental Framework

2.1.1: Transcendental Idealism

Kant's Transcendental Idealism is sometimes referred to as an epistemological idealism, distinct from the ontological idealism defended by philosophers such as Berkeley. Where ontological idealists generally defend the view that reality is constituted of the mind, and no mind-independent reality exists, Kant takes an ambivalent position on this issue. Kant argues for looking inward, at the structure that determines the way that we experience the world, and limiting our epistemological claims to that domain only[11].

Kant can be said to start from a point of indirect realism, recognizing the appearances of objects need not reflect their full nature, with our cognitive system acting as a mediator and rendering a consistent image based on transcendental (inherent) structures. He expands this claim to include his famous distinction between Noumena and Phenomena, where Noumena references things in themselves, i.e. in the real world, and Phenomena are the resulting experiences that we have access to and experience as our life-world[12].

2.1.1: Noumena and Phenomena

Kant's Noumena are unknowable to us. We can infer that *something* is causing Phenomena that we experience, but cannot say anything about what shape or qualities these objects may inherently possess, due to the necessary cognitive filter that they have to go through, limiting any epistemological claims we can posit about their epistemological status. In this sense he takes the indirect realist position and adds to it an idealist twist; he does not argue that Noumena are mind-dependent, but also doesn't position for the opposite. As we cannot know Noumena, Kant says, we should focus on the world as we can experience it in Phenomena, as any sort of knowledge we will be able to generate is only valid towards Phenomena. As such, Kant doesn't necessarily imply that reality is mind-dependent, as ontological idealists do, but rather that the things we may know are mind-dependent, making his transcendental idealism an epistemological idealist stance.

What's important to remember here, as will be discussed again in the section on LLM Transformer models, is that the experience we have of the world is thus by necessity limited to a narrowed down version as structured by our cognition. As we will see at that point, a similar thing can be said for these LLM models. Thus, if we can take inspiration from the specifics of Kant's framework and find equivalencies in the model architecture, we may generate new insight into their limits and possibilities for future development.

---

[11] Stang, Nicholas F. "Kant's Transcendental Idealism." *In The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2024 Edition. https://plato.stanford.edu/archives/spr2024/entries/kant-transcendental-idealism/

[12] Ibid.

2.2: The Transcendental Pipeline

2.2.1: Raw Data, Sensibility, and Intuition

In order to draw comparisons between the systems of Kant and Transformers, we first need to get a good overview of what raw sense data goes through in Kant's system such that we have a comparison point for Transformers later. Thus we will try to create a high-level overview of the Kant's system and the most important components that make knowledge generation possible. We will start from the raw sense data and provide a pipeline that goes all the way to coherent experiences and knowledge forming in Kant.

For Kant, we begin with a raw sensation, i.e. any sort of signal we may pick up whether it fits into our experience or not. The first step this manifold of raw sensations take in our pipeline is through the faculty of sensibility. Here, the sensations are ordered through the pure intuition, which for Kant is the form which all data has to take for it to be interpretable; namely in the form of space and time. Thus, any sensation that we have is by necessity ordered in space (dimensions) and in time, as basic building blocks of our cognition. It is also because of this that a priori synthetic statements, such as those concerning geometry, are possible, as the mind has space built into its very structure. This sensation captured in the forms of the pure intuition is referred to as an empirical intuition[13].

2.2.2: The Imagination and The First Syntheses

We can then move on to the faculties of understanding and imagination, which roughly function as that which integrates sensory experience into a coherent whole, and that which applies a priori constructs or categories to it. This is a bit of an simplification as these faculties can only work in unison, and the imagination is involved in more than just integrating sensory experience as it also helps the understanding apply itself as we will see in the schematism. Still, this divide is useful for explaining the faculties separately first. We start by spending some time on the faculty of the imagination.

Once the empirical intuition has been formed, it is still nothing but a flurry of unstructured information, given to us in space and time. In order for it to be interpretable, we need to combine these empirical intuitions, or manifold, into a coherent whole. This is done with a three-fold synthesis which encompasses all 3 major faculties as they have been introduced so far. This process starts with a double synthesis with sensibility and imagination on the manifold to structure it more clearly: The Synthesis of Apprehension, and the Synthesis of Reproduction, take place in unison as structured experience cannot be created without both working in tandem[14].

The Synthesis of Apprehension has the imagination take the manifold of intuition and order them in space and time[15]. Note that this is different from the pure intuition, in that where the pure intuition only *delivers* sensibility in space and time, the Synthesis of Apprehension *orders* the manifold in these dimensions. This is only made possible when the mind can also keep track of which objects of experience change and which remain the same from any single moment to the next, meaning that these objects need to be reproduced at the same time, as an ordering without recognition would still be uninterpretable. The Synthesis of Reproduction allows this recalling of sensations and relations in the past, reproducing them, so that a total representation can be made of the object we are grasping[16]. In this sense, the imagination can be said to have a reproductive and a productive function, in that it allows us to both recall and then amend or produce the representations that we have.

---

[13] McLear, Colin. "Kant: Philosophy of Mind | Internet Encyclopedia of Philosophy," n.d. https://iep.utm.edu/kantmind/#SSH1ai

[14] Brook, Andrew, and Julian Wuerth. "Kant's View of the Mind and Consciousness of Self." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2023 Edition. https://plato.stanford.edu/archives/spr2023/entries/kant-mind/

[15] Kant, *Critique of Pure Reason*, A99

[16] Kant, *Critique of Pure Reason*, A100

2.2.3: The Understanding, The Last Synthesis, and The Unity of Apperception

Now that the imagination has produced a steady of the objects for us in which its materials are constructed temporally and spatially, we need to be able to relate the object to the past in order to truly be able to take up the object into our cognition and gain knowledge about it. This is necessary as we otherwise wouldn't be able to understand that the same object is being represented by differing representations over time. This is where the faculty of understanding comes in with the last of our three-fold synthesis in the Synthesis of Recognition.

Whereas the previous syntheses worked on ordering the empirical intuitions from a manifold into a coherent whole, this last synthesis, as requiring relation between an object over time, needs systematic rules to tie them together. That is to say that it needs memory, and relations of that object[17]. These rules or concepts can be found in the categories of the understanding: quality, quantity, modality, and relation. Relation specifically is not yet necessary here as we are still talking about a single object given in experience, and will be expanded on soon. Of course, as these categories are a priori and thus not directly applicable to empirical intuition or the manifold thereof. To apply them to the empirical intuitions indirectly, Kant introduces the 'schematism' in which the imagination applies schemata to each of the 12 subcategories of the understanding in a 1:1 ratio that do the mediating between these pure concepts and the intuitions[18]. Then finally, the product of this is united into a stream of consciousness of sorts in the unity of apperception.

This is where relation comes in. Kant argues that in order for different representations of objects to exist at the same time rather than only one over time, we need to be able to relate these different representations to one another. This is really where the focus has to shift from discussing the processing of the empirical intuition to the transcendental subject itself. After all, someone has to be the one relating these different representations to one another. For this, Kant argues, we need the unity of apperception, which is to say the combining of all of these different representations over time and space into the singular experience of an *I*. Our minds need to be able to do this a priori as this *I* is not inherent in the empirical intuition, which is why the transcendental subject has the faculty of the unity of apperception, which does this very task[19]. Thus, this unity of apperception needs to allow us to be cognizant of our synthesizing, and gives us the ability to recognize that *I*-as-being am doing this synthesizing, it is thus the stream of consciousness that binds everything together.

[17] Kant, *Critique of Pure Reason*, A103
[18] Kant, *Critique of Pure Reason*, A140
[19] Kant, *Critique of Pure Reason*, A115

2.2.4: Judgement and Reason

So now we have sensory data, being filtered by the sensibility and the imagination through the use of categories of the understanding, and everything being brought together into a singular consciousness by the transcendental unity of apperception. The final step is to take this stream of experience and create knowledge, that is to say to judge what comes in and form frameworks that capture these experiences. This is the function of the pure faculties of judgement and reason. Judgement in particular takes the experience of the manifold of intuition as held together by the unity of apperception to a conceptual level, where one may recognize that the object as, for example, a cup of coffee, rather than a stream of experience (blackness, bitterness, etc.)[20]. This is crucial for gaining knowledge, as it allows for the subsuming of particulars into generals, forming the basis of reason to then operate on. The faculty of reason finally takes these concepts and reasons with them, relating higher level concepts to one another, as one then realizes coffee is bitter because of a particular type of bean, for example, and can be sweetened with the use of sugar. We have now reached knowledge generation, and thus intelligence. The next step is to analyze the Transformer architecture in a similar way, before contrasting their systems and figuring out what implications this will have.

3: The Transformer Architecture

The transformer architecture, first published in a now famous 2017 Google research paper titled "Attention is All You Need"[21], forms the bedrock of the new wave of AI applications. As we have seen with Kant's system, a complex interplay of a priori faculties is needed for the functionality of subjective (in the transcendental sense) knowledge. The specific counterparts of these faculties in AI systems will depend heavily on the specifics of the system we are examining, as even with a baseline-model such as the Transformer Architecture, specific models will have different interpretations and differences according to their own design needs and capabilities. As such, our analysis of these components will stick a bit more to the basics that are applicable to all such models, and will be mostly explained through the specific example of GPT (OpenAI), due to its popularity not only as the most used LLM-tool, but also the corresponding availability of information regarding its architecture.

3.1: First Things First: A Quick Overview of AI-Systems

For those unfamiliar with Transformer architecture in particular or AI in general, we will quickly dive into where such an architecture is positioned within the AI hierarchy. AI is a general-use term referring to systems that mimic human-intelligence, i.e. mimic intelligence artificially. Depending on who is asked, this can range from simple trees of if-then statements given an input, all the way to modern generative models. It is here that we can already make a distinction between the two general horizontal approaches mentioned in the introduction in Symbolic AI and Connectionist AI[22].

---

[20] Hanna, Robert. "Kant's Theory of Judgment." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2022 Edition. https://plato.stanford.edu/archives/spr2022/entries/kant-judgment/

[21] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." arXiv.org, June 12, 2017. https://arxiv.org/abs/1706.03762

[22] Goel, Ashok K. "Looking Back, Looking Ahead: Symbolic Versus Connectionist AI."

3.1.1: Symbolic AI

Symbolic AI is the general model approach focused on the encoding of rules into the system that it may then follow to create decisions, like one would do in standard software engineering. This rules-based approach greatly increases the explainability of model decisions, as the steps it follows to manipulate data come from clearly defined instructions, and can thus be traced rather easily[23]. This falls roughly in line with Rationalist epistemology in Philosophy, where the faculties of the mind operate on information to synthesize new insights according to clear steps, thus creating knowledge. The drawback of these models is that they don't deal with unstructured data very well, as the system is not really able to 'learn' or deduce the underlying patterns in the data, given that its rules are set in stone.

3.1.2: Connectionist AI

As Connectionist AI is the approach that has given us Transformer models, we will begin with some wider context and the sub-section of AI that it falls under, being Machine Learning (ML). ML systems allow the model to learn from the data; i.e. the system allows the data to determine the parameters of the model. While we will elaborate on parameters more later, the main idea is that these are the variables in the model that will be trained on data, thus allowing the model to create its own network of variables that work well for processing and labelling the data it has been given.

More specifically, Connectionist AI approaches fall under a sub-section of ML called Deep Learning (DL), in which many different individual nodes are interconnected to capture more complex patterns than a traditional ML-model would. It is, as the name implies, based on the human brain and the connectedness of its neurons, which themselves are said to work by strengthening or weakening links between one another. This effectively means that the nodes (neurons) and weights (synapses) are calibrated based on the data (input) that they receive. This means that the 'rules' of the system for processing data are determined by only the input data (external stimuli) and hyperparameters, which are the human-set parameters that determine the conditions under which the model may tweak its parameters. Thus, for any given hyperparamatized system, the eventual output of the system is solely dependent on input data, in line with the Empiricist approach. This also gives us an initial point of contention for our later analysis, as we can already see from this setup alone that the lack of transcendental- or rule-based parts of Kant's system may form a problem for intelligence in these models.

We can now look at Transformers, which are a type of DL-model based on the principle of Attention. From a high-level overview, Transformer models will consist of many layers of Attention- and Multilayer-Perceptron (MLP) blocks, which are alternated between[24], before a prediction is made of the next object (token) in the sequence (text in GPT's case). As MLPs form the bedrock of the general Connectionist approach and can provide us with the best basic understanding of how Deep Learning models set their parameters, we will start at the MLP-layer. We will then analyze the Attention layer as well, before tying it together by combining the Transformer functions to Kant's framework in the next section.

3.2: Diving Into the Details

As mentioned before, the transformer architecture can be broadly categorized as having 2 main layers, the Attention layer and the MLP layer. We will start by spending some attention on the technical details of MLPs, so that the general intuition on how model parameters are set is built up.

---

[23] Varone, Marco. "The Best Part of Symbolic AI: Full Explainability." expert.ai, May 27, 2022. https://www.expert.ai/blog/the-best-part-of-symbolic-ai-full-explainability/.

[24] Ashish Vaswani, et al., "Attention Is All You Need."

3.2.1: Multilayer-Perceptrons and Basic Deep-Learning Architecture

A MLP is a network that has an input layer, one or more hidden layers, and an output layer. Each of these layers, other than the input and output layer, can be thought of as 1 dimensional vectors which are connected to the layer before and after them with so-called weights. For the Transformer architecture, a Feed-Forward Neural Network is used, meaning that these inputs only travel one-way, and auto-regression is realized. More attention will be given to this later. The network works by giving it an input in the form of a vector, which is then propagated through the layers. Each individual node after the input layer will be 'activated' by the sum of product of the connected neurons' activation with the weight of the connection, minus a bias, and normalized. Assuming all nodes between layers are connected, that gives us:

$$a_j^L = ReLU(\sum_{k=0}^{n_{L-1}} (a_k^{L-1} \cdot W_{jk}) - b_j^L)$$

Where:

a_j^L is the activation of neuron j in layer L

W_{jk} is the weight between neuron k and neuron j and a model parameter

b_j^L is the bias term of neuron j in layer L and a model parameter

$ReLU\ (x)$ is a function such that $ReLU\ (x) = Max\ (0, x)$ and a model hyperparameter

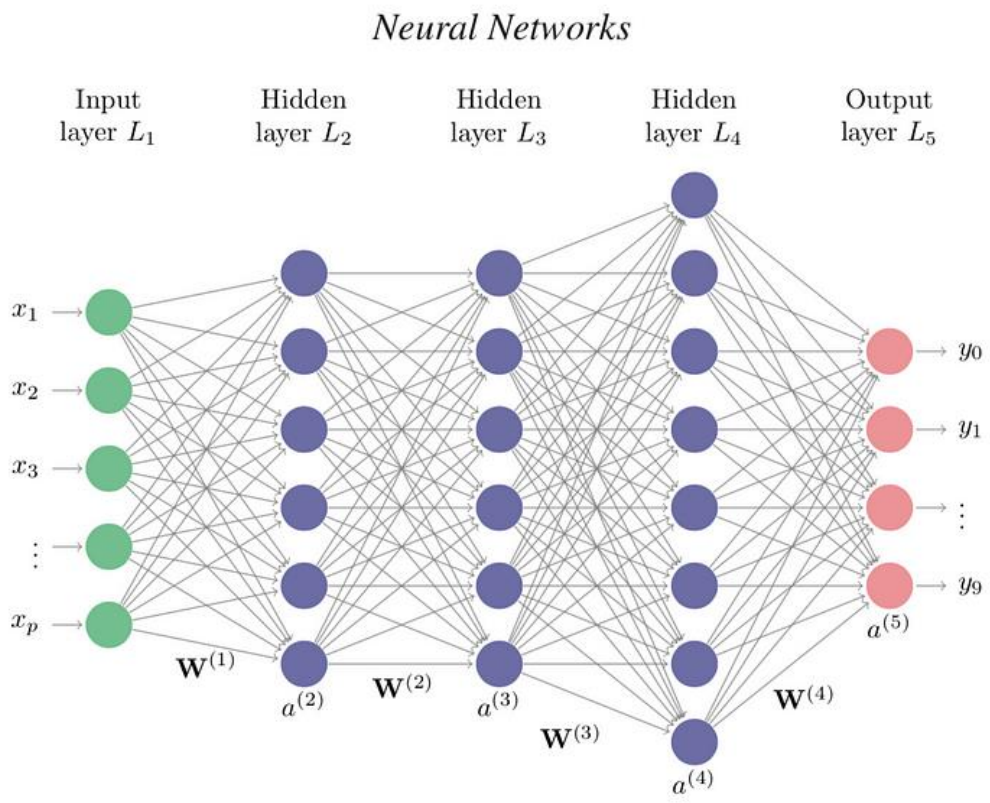

Figure 1: Visual Aid for MLPs[25]. Credit to Kose, Parvez.

[25] Kose, Parvez. "Feed Forward Neural Network — Explainable AI Visualization (Part 6)." *Medium*, February 20, 2023. https://medium.com/deepviz/explainable-ai-and-visual-interpretability-background-part-6-6467736f82b8

This may look complicated, but makes sense if we remember that these systems are modelled after the brain, with its neurons and synapses. In simpler language, any node's activation is just decided by how activated the nodes in the previous layer are, and how strong the connection (synapses) is to them (see figure 1 for visual aid). Furthermore, the specific choice of activation function (ReLU) is arbitrary for the purposes of this thesis as it is often utilized over alternatives such as Tanh and Sigmoid because of its efficiency in compute and effectiveness against the vanishing gradient problem[26], which are irrelevant details in our epistemological context. For us, the important detail for the activation function is that it cuts out values under 0 and thus introduces non-linearity in the system, enabling the model to capture complex patterns more effectively.

Keen readers may notice that the activation function does not include any way of averaging over the sum of its connections, meaning an activation could become arbitrarily large rather than capping at 1. This is where the bias-term comes in, which is another model-trained parameter. This means that as the model can set this based on data to balance strong connections, extra degrees of freedom are added to the model, adding to the complexity and ability to capture patterns over simply averaging out the values in the sum. As such, the system will learn over time by adjusting the weights and biases, comparing model performance to the desired, labelled, training data. It does so having a so-called Cost Function, which takes the difference between the desired and actual output of the model over the whole output layer (as opposed to the single-node activation function above), which looks as follows:

$$C_0 = \sum_{j=0}^{n_{L-1}} (a_j^L - y_j)^2$$

Where:

C_0 is the total cost of the entire output layer for one given training example

a_j^L is the activation of neuron j in layer L

y_j is the desired result (activation) of neuron j

[26] Wang, Chi-Feng. "The Vanishing Gradient Problem - Towards Data Science." *Medium*, December 7, 2021. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

3.2.2: Learning Through Gradient Descent and Backpropagation

In order to minimize this cost function, Gradient Descent is used. Gradient descent is an optimization algorithm that tries to find the minimum of the cost function through iteratively changing the parameters of the model such that the Cost decreases. In order to figure out which parameters to change and in which direction, the most common used efficient algorithm is called Backpropagation.

Backpropagation takes the partial derivatives of the cost function with respect to the model parameters, and makes a small step in the opposite direction. That is to say that it effectively looks for small steps it can make in a_j^L such that the cost function decreases until it finds a minimum in which the derivate is equal to zero. Readers familiar with modelling may know that a model can have multiple minima in their functions. By employing Backpropagation, the algorithm can stop at any given local minimum, as the gradient will be equal to 0 at that point[27]. This means that while multiple local minima may be found by utilizing different starting points, no guarantee is given that the true minimum of the cost function will be reached. This consideration is part of what informs the step size of the Backpropagation algorithm, which determines how by how big of a step model parameters can be changed, where a larger step is more efficient and allows the algorithm to jump from one downwards slope to the other (by overshooting the minimum). As such, the hyperparameters for step-size will usually make the step-size dependent on the slope at any given point, making the steps progressively smaller as to not overshoot the local minimum (see figure 2 for visual aid).

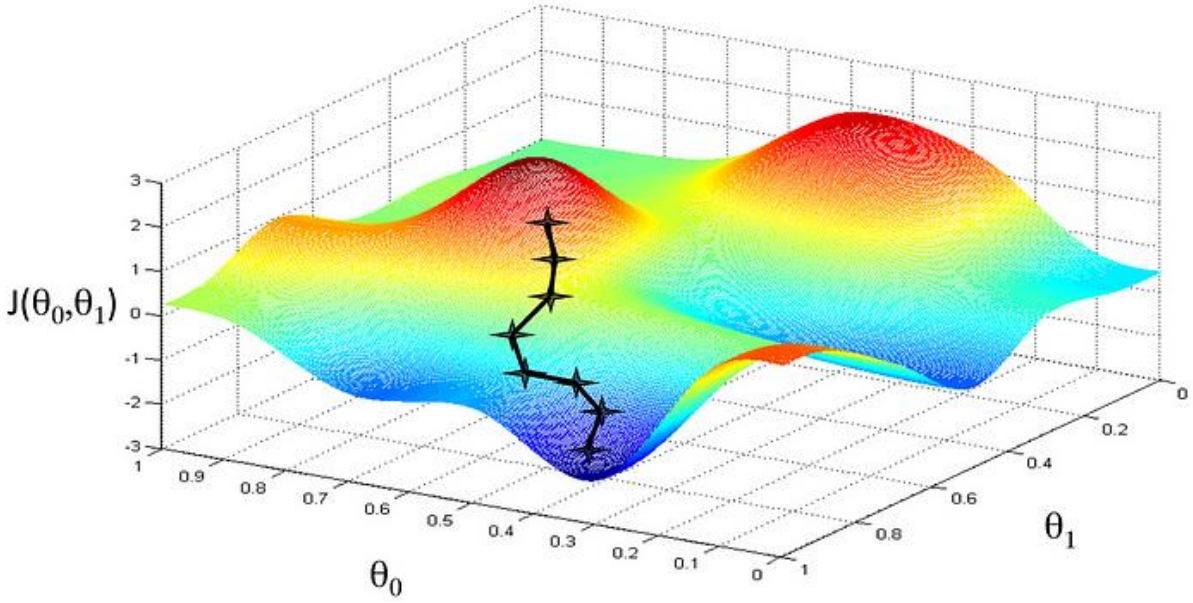

Figure 2: Visual Aid Gradient Descent[28]. Credit to Jaleel, Adejumo.

[27] Zhang, Jiawei. "Gradient Descent Based Optimization Algorithms for Deep Learning Models Training." arXiv.org, March 11, 2019. https://arxiv.org/abs/1903.03614

[28] Adejumo, Jaleel. "Gradient Descent From Scratch- Batch Gradient Descent, Stochastic Gradient Descent, and Mini-Batch Gradient Descent." *Medium*, April 16, 2023. https://medium.com/@jaleeladejumo/gradient-descent-from-scratch-batch-gradient-descent-stochastic-gradient-descent-and-mini-batch-def681187473

Looking into the calculus of backpropagation is not really useful for our purposes, as it simply involves showing how the partial derivates of parameters decrease the cost function. Rather, the main function for us to keep in mind is that single training examples are used to compare the output to the real label to compute the cost function, after which the backpropagation algorithm integrates this example by adjusting our parameters to decrease the loss function towards a local minimum. Thus, the model 'learns' by adjusting the parameters for each training example such that the model should eventually figure out what patterns to look for in a new example such that it provides the right output. The training phase is finished when the model creator deems that the model performs well enough on test data, i.e. data is has not seen before, to release it.

A final important caveat here is that models will sometimes overfit to the data, which means that the parameters are so accurately trained on the training data and given so many parameters to tweak to achieve this goal, that generalization into testing data actually starts to suffer. This means the loss will start to increase beyond a certain amount of training and parameter-space, as the model learns to capture not only the underlying patters in the data, but also the noise[29]. There are ways around this, such as dropout, which deactivates neurons during training, or regularization, which adds a penalty term to the loss function to disincentivize strong weights. This notion of overfitting will also become important for our epistemological implications as we will discuss later.

3.3: The Attention Mechanism

Now that we have an idea on how these DL-models learn, we can see how they deal with new data and knowledge. For this, we will turn our attention to Attention. The transformer model is at the forefront of the recent AI-boom, with its trump card being the Attention mechanism. Let us run through it using text as an example in line with LLMs and OpenAI's GPT.

3.3.1: Embeddings and Vectors

Starting from a similar point as in our Kantian analysis, we will look at embedding first. Embedding is the translation from an index into a vector. The process that takes place starting from a corpus of words to the model learning about their meaning looks roughly like the following: First, the words go through an algorithm that tokenizes them; i.e. splits them up into full words, smaller parts, or even specific characters, based on the specifics of that algorithm. These tokens are then indexed, i.e. a model vocabulary is made and each token is given a specific identity that the model can recognize. Then, from this index, embedding turns the identities into vectors in a high-dimensional space[30].

The specifics of this process are non-arbitrary, as they will have a great effect on the way the model is able to learn. Specifically, simple older models may use one-hot-embedding, which pretty much just leaves every index as its own identity. This works if you simply want to have words living in separate spaces from one another, but does not lend itself well for the model to learn the meanings that they bring, due to the overfit not allowing embeddings to relate to one another[31]. Rather, by limiting the dimensionality of the model and not allowing each token its own space, the model has to learn during training how close certain words are to each other as they occupy a similar spot in the high-dimensional space, which is ultimately where their initial meaning before the imbuing of further context comes from.

---

[29] Ying, Xue. "An Overview of Overfitting and Its Solutions." *Journal of Physics. Conference Series* 1168 (February 1, 2019): 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

[30] Dahouda, M. K., and I. Joe. "A Deep-Learned Embedding Technique for Categorical Features Encoding." *IEEE Access* 9 (2021): 114381-114391. https://doi.org/10.1109/ACCESS.2021.3104357

[31] Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 16 (April 2018): E3635–E3644. https://doi.org/10.1073/pnas.1720347115

To posit a widely used example for this: The dimensional difference vector between the tokens for "aunt" and "uncle" is likely to be very similar to the difference between "daughter" and "son", encoding a rough gender vector. Still, the (aunt-uncle) and (daughter-son) difference will have more subtleties from daily use, as can be demonstrated by (queen-king), where the difference will be bigger due to other contexts of token-use such as the Queen rock band. What's important to remember here is that we are dealing with a very high-dimensional-space, meaning that these meanings go much deeper than gender[32]. GPT-3 had about a 50 thousand token vocabulary, while GPT-4 doubles that, according to various unofficial sources. The first layer of the transformer then comes down to taking in the input, and translating it into vectors, essentially just looking those vectors up from the vocabulary.

3.3.2: The Heart of the Model, Attention-Heads

The heart of the model, the Attention mechanism, will now take all of these initial tokens with their initial meanings (values in the vector), and have them influence each other to capture complex meaning. After all, a table has a different meaning in the sentence "I left some milk on the table" than in the sentence "Please check the excel table for the updated values". The attention mechanism seeks to capture these complex relationships. To understand it, it is good to keep in mind that a Transformer model never works with one attention head, but rather a large number running in parallel. The way to view this conceptually is that one head of attention will focus on one type of meaning for our input, like syntactic relationships, while another attention head may focus on something completely different like the semantic relationships. These heads and the exact meanings they are trying to capture are as we will see emergent properties from the parameters of the model and are thus hard to pin down. For simplicity's sake, we can imagine that our example sentence is made up of only full-word tokens as opposed to smaller sub-parts of words, and our attention-head a focusses on the relationship between nouns and adjectives.

3.3.2.1: Key- and Query Matrices and Vectors

The input into the model will be embedded not only in the meaning of the word from the vocabulary, but also its position in the input sentence(s). Then, for each token (in this case word), they are multiplied by 2 matrices in the form of a Query and Key matrix to create a Query and Key embedding. These matrices are again part of the parameters of the model, and differ by the specific attention head. The way to think of this conceptually is that the Query embedding is the asking a question for its associated token embedding, and the Key embedding giving an answer to that question. In our example of nouns and adjectives, the nouns in the sentence may be asking "Do I have an adjective in front of me?" where the Query embeddings for the adjective will be answering "Yes, I am" in this case made possible by the positional encoding of the embedding, as well as the adjectives and nouns being encoded as such in their embedding vectors as described in the previous section (with the gender example)[33].

The Query- and Key vectors of every embedding are then multiplied using the dot product, which means that when the vectors are similar, the value of the dot product will be bigger. These values are then normalized for each individual token, using a function called Softmax. What this means in practice is that for a given text, all the other tokens that are relevant to the specific token according to their dot product, will have a value greater than 0, with the total adding up to 1, thus creating a sort of distribution of relevance of the Key-associated token to the Query-associated token. If a token is relevant to another, this is what is referred to as 'attending to' which is the core of the model. The total mapping of how each token attends to another is called the Attention Pattern, which again comes down to relevance of tokens on others.

---

[32] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." arXiv.org, January 16, 2013. https://arxiv.org/abs/1301.3781

[33] Ashish Vaswani, et al., "Attention Is All You Need."

3.3.2.2: Value Matrices and Vectors

Now that we have the Attention pattern showing us how *relevant* each token is to each other token, we now need to figure out the *value* each token provides to other tokens. For this, yet another parameter matrix is used, called the value matrix, which again combines with the token embedding to form a value vector or value embedding. Conceptually, we can think of the attention pattern as showing how relevant a token is to another on a scale from 0 to 1, and the value vector as showing in what way the token is relevant to other tokens. Essentially, in line with our adjective-noun example, "a red car next to the store" would have "red" attend to mostly "car" and not to the other tokens, while the value vector would then update "car" with redness, while "store" is not updated as the attention value would be close to 0.

3.3.2.3: Adding it All Together

The logical next step is summing up the updates that any particular token receives from all the others into a single change vector. This vector can then be added to the original token, and we now have a token that has soaked in all of the meaning from the others in whatever dimension that particular attention head was looking for. In our example, nouns now have the meaning of adjectives imbued in them, like the car is now updated to contain redness, like in the (aunt-uncle) example at the start. The attention mechanism is summed up as:

$$Attention\ (Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where:

QK^T is the table of the products of the Query and Key dot products

V is the value vector for each token

$\sqrt{d_k}$ is the square root of the dimensionality of the Key-Query space.

$Softmax$ is the normalizing function that turns a vector of real values into a vector of real values that sum to one. This is utilized on the columns in the Key-Query space, i.e. the Keys.

3.3.3: The Broader Picture

The full transformer architecture alternates between Attention and MLP blocks as described above. This allows the model to have extra dimensionality and thus capture complex meaning. As the MLP block will have as input the updates tokens from all separate attention heads, it will find the patterns in how these different ways of looking at meaning can combine and influence each other, before outputting the now updated tokens through the next layer of Attention to try and capture more nuanced meaning. This can continue for many layers depending on the model's hyperparameters.

Finally, at the last layer, the last token in the input, imbued with all the meaning from all of the different layers, is passed through the projection layer[34], which is yet again a parameterized matrix and bias that is learnt through the training phase. This projection layer brings the token from a meaning imbued embedding to a vector with the same dimensionality as the initial vocabulary vector, where the values of the vector are the logits corresponding to words in the vocabulary vector. By applying Softmax here, we are left with a vector that has a probability distribution of possible next tokens, which will be used to produce the next token in the sequence, i.e. word in the text.

[34] Ibid.

How the next token is chosen depends on the model setup. This can be done by simply picking from the probability distribution, but can also be altered by using for example temperature-scaled sampling, such as GPT does. In that case, a temperature of 0 means the model will always pick the token with the highest probability, making the model deterministic, whereas a higher temperature will pull the probabilities closer together, creating a more stochastic model.

We now have a good understanding of the way Transformers transform input from the training phase into a model that can combine nuanced meanings and come to surprising insights from a user's perspective. Whether this holds up to the scrutiny of true intelligence, however, remains to be seen. Towards this end, we will now contrast this architecture with Kant's Transcendental system to figure out what implications arise for the models and their future development, as well as policy making and epistemology.

4.1: Synthesizing Our Systems

We now have an understanding of the necessary conditions for human cognition according to Kant, as well as the conditions under which Transformers manipulate their parameters for synthesizing information. The next step is to contrast the two systems in order to find any structural differences that could give us insight into the nature of intelligence and possible model development. In line with the previous analyses, we will be starting from the input and working our way towards the output or knowledge generation.

4.1.1:Raw Data and Its Dimensionality

Starting from Kant's notion of sensibility, we might recall that it is used to structure raw data into a specific format that we have access to through the pure intuitions of space and time. That is to say that any data that we get to process is necessarily packaged in these dimensions of space and time. This can be said to resemble the way that Transformer models can only take vectors of real numbers as input to create their representation of reality pretty well. In both cases, possible parts of reality, or 'things in themselves' are lost with this conversion. In Kant's system we refer to this faculty as the sensibility, whereas for Transformer models, this happens with the tokenization of reality. Note that we are not talking about embedding yet, as that already has some sort of meaning or relations attached to the tokens. We are thus left with empirical intuitions for Kant, and data arrays for our ML-model. As this equivalence seems to hold pretty well, we can move on to the processing of this dimensionalized data.

4.1.2: The Synthesis of Apprehension and Reproduction: Recreated in the Aggregate?

The next step we discussed in Kant's framework is the threefold-synthesis, starting with the combination of the Synthesis of Apprehension and the Synthesis of Reproduction to order the empirical intuition in time and recall the past state of their respective objects. While these appear to have no direct equivalent in our model, an argument could be made that the Synthesis of Apprehension is in some capacity replicated by the positional encoding of the tokens. While we used a text example without continuous time-flow, it seems a reasonable leap to imagine that in different input modes such as video, this could be replicated. Each frame that is fed into the model might have a time-encoding, for instance, effectively ordering the experience through time as the Synthesis of Apprehension does.

The Synthesis of Reproduction has no such analogous counterpart, as Transformer models don't have an ingrained, underlying concept of an object that they could reproduce. We could feed the model completely separate, unrelated data with subsequent positional encodings, and the model would try to make the best of it and end up with a random model-structure, not realizing its error. Humans wouldn't run into this problem, due to their self-apperception. However, I will later argue that in a high enough dimensionality, Transformer Models could imbue their tokens with an emergent dimension for object-ness. Still, this is not at the root like with Kant; more on this later.

4.1.3: The Synthesis of Reproduction and the Status Problem

This notion of object-ness and the relations between them becomes important when we start to look at the application of the faculty of understanding through the Synthesis of Recognition. As discussed prior, the Synthesis of Recognition requires memory and the relation of an object with itself over time, as we cannot cognize an object properly without understanding how a previous state may lead to a later state. For this, the categories of the understanding are used in Kant, which poses a major challenge for the Transformer architecture.

Transformer models work by mapping the meaning of embeddings in a high-dimensional space before updating them based on the surrounding context using Attention. Whereas Kant's system has set-in-stone categories that underlie experience mediating this updating and take into account previous object states, embeddings have no such thing. This means that this previous state has no special status, and thus no epistemic discernability would exist between an embedding's past states as expressed in the current, and an arbitrary part of the vector. Kant's categories are turned into just another inference made on the basis of experience (training). This remains true if we stop looking at the individual embedding and instead focus on another way in which the model can capture relations in the weights and biases of the MLP-layer. Here, too, we can imagine that the model would be able to capture any of the categories as patterns of nodes and weights, but the same problem of epistemic status persists.

This status problem may seem trivial, but is far from it. After all, these models learn their dimensions on the basis of inference on data, and function practically as a black-box due to the complexity of the system. Thus, in real-world scenarios, we would have no guarantee that these concepts would be fully captured in the embeddings or model weights at all, except in idealized cases as will be discussed later. This effectively means that we have no guarantee Kant's concepts such as causation will be fully present in the model's life-world.

This has further ramifications for the entire epistemic set-up of the model. For one, with the categories gone outside of idealized cases, Kant's schematism is no longer necessary as there are no rules and intuition to mediate between, and only tokens are used for the updating of other tokens by imbuing them with context (i.e. relations). The same problem of epistemic status in real-world contexts still remains, however, which means that no guarantee can ever be given on the model's effectiveness in capturing the world adequately given its black-box nature.

It is with this insight that a new split occurs that hadn't seemed completely relevant before: Rather than simply looking at Transformer models in either a near-future state or an idealized state in which compute and data are infinite, a third classification seems to be relevant: a state of very exponential compute and data, without invoking (close to) infinity. This distinction seems nonsensical at first, but arises when trying to cognize the ways in which the model may imitate Kant's system. We can namely posit that while exponential compute won't fix this status problem, infinite compute paired with perfect training data might. Let's analyze each state in more detail.

4.1.4: The Status Problem Tackled in Different Model States

4.1.4.1: The First State: Expected Short Term Model Development

In this state, we can imagine the models are they are now, with slightly added compute and training, mirroring development as we may expect it to happen over the coming months. Here, we quickly run into the limitations of the architecture: No equivalency to Kant's faculty of Understanding can be drawn, which means that while tokens can be updated based on context, real reasoning from first principles (using the categories of the understanding) cannot take place. This means that model is likely to be able to only create 'knowledge' as long as it falls within the confines of the available data and patterns that it has been trained on. It may be able to update itself on the fact that the embedding of "chair" has some sort of direction in its vector that points to it often having the quality of "wood", meaning that in the right context, e.g. when talking about a fire with the quality of "burning", it could conceivably imbue the chair with that meaning. Still, whether or not the model would accurately capture those specific patterns is impossible to figure out in practice, given the status problem.

In a real life scenario, that would also mean that the model is vulnerable to attacks that exploit its pattern recognition. An example of this is the use of Generative Adversarial Networks (GANs) in which 2 models play a zero-sum game. The objective is as follows: each model has to trick the other into making the wrong prediction based on the input data that it is given. Thus, a model is generated and trained with the sole purpose of fooling another model, roughly similar to Gettier cases in Philosophy, where specific examples are presented to prove the triparte-account of knowledge doesn't capture all necessary conditions. In the context of GANs. this can happen by, for example, creating small perturbations in an image which throw off the pattern recognition of the model while remaining invisible to the human eye. This not only highlights the model's dependence on robust data, but also the risks and unreliability that such a model will always bring in real-world scenarios if 100% accuracy is the goal.

4.1.4.2: The Second State: Near Infinite Compute, Laws of Physics Hold

This state needs a little bit more explaining to distinguish it from the third state later on. By exponential but not infinite compute, we are referring to any arbitrarily large number of parameters and data that don't explicitly break the laws of physics. The distinction here can be easily explained with a game of chess.

In a game of chess, a very large amount of combinations of game states and flows are possible. So much so in fact, that Shannon argues that there are exponentially more possible games of chess to be played than there are atoms in the observable universe[35]. This also implies that keeping a full database of all possible chess games, such that you can "solve" the game mathematically (i.e. always look up the guaranteed winning move from a lookup table), would be impossible, as there simply are not enough atoms in the universe to even store that information in the first place. We distinguish between this state and state three similarly.

The limitations that come from state one are alleviated in this state. With enough data and compute, the model may reach any arbitrary amount of accuracy in capturing patterns and thus being able to 'predict' (through the attention mechanism) following events. This would, in the extreme, mean that the model could be completely intelligent from a functionalist perspective, in which the 'reasoning abilities' of Kant's system are captured so well by inference that the embedding updating would be virtually indistinguishable from human or even superhuman reason and knowledge creation.

---

[35] Shannon, Claude E. and Bell Telephone Laboratories, Inc. "Programming a Computer for Playing Chess." *Philosophical Magazine*. Vol. Vol. 41, March 1950.
https://vision.unipv.it/IA1/ProgrammingaComputerforPlayingChess.pdf

From an epistemological perspective, however, we would still never be able to know how the reasoning works, and whether it is correct and fully captures everything, simply because of the black box nature of the model, and the fact that we know it cannot infer literally every aspect and relation due to the chess-analogy above, retaining the status problem. This also implies that while the robustness of such a model may increase, it would still be vulnerable to new or distorted data, such as those generated by GANs as seen above. Still, it would likely, with extreme improvement of the data and parameters, be functionally equivalent to human reason, simply by virtue of its high accuracy and humans being unable to tell the difference.

4.1.4.3: The Third State: Truly Infinite Compute and Perfect Data

When given infinite compute and perfect training data, something interesting happens to the model: it becomes Turing Complete, and thus starts to be able to reason[36]. This means that in an idealized world with this state, we can now fully trust on the Transformer to be able to reason using inference alone, getting rid, -from a functionalist perspective-, of the constraints imposed by Kant's categories of the understanding. This occurs as in this idealized state, any arbitrarily small sub-part can have its own token with unlimited context. Any token representing an object could thus have imbued within them with certainty that they are beholden to any specific law of physics, for example. Given the Universal Approximation Theorem[37], which states that a model can approximate a continuous function until an arbitrary accuracy given enough compute, we could capture functionally everything given the perfect data and infinite computing power.

This does not mean, however, that this model is limitless, but rather that it runs into the same limits that the blocks it is built on runs into; i.e. it runs into the limitations of math itself. Being a math-based machine, it ultimately runs into the problems of completeness, consistency, and decidability as posited by Gödel's two incompleteness-theorems[38] and Turing's halting problem[39]. What effects this may further have is ultimately outside of the scope of this thesis, however.

---

[36] Stogin, John, Ankur Mali, and C. Lee Giles. "A Provably Stable Neural Network Turing Machine With Finite Precision and Time." *Information Sciences* 658 (February 1, 2024): 120034. https://doi.org/10.1016/j.ins.2023.120034.

[37] Lu, Yulong, and Jianfeng Lu. "A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions." Conference-proceeding. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, n.d. https://proceedings.neurips.cc/paper/2020/file/2000f6325dfc4fc3201fc45ed01c7a5d-Paper.pdf

[38] Raatikainen, Panu. "Gödel's Incompleteness Theorems." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2022 Edition. https://plato.stanford.edu/archives/spr2022/entries/goedel-incompleteness/

[39] Copeland, B. Jack. "The Church-Turing Thesis." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2024 Edition. https://plato.stanford.edu/archives/spr2024/entries/church-turing/

4.1.5: The Rest of the Pipeline

Now that we have seen how the lack of categories of the understanding impact model knowledge creation or intelligence in different model states, we can move on with the rest of the Kantian pipeline. The next factor in this is the Unity of Apperception, and the lack of an identity that the model has for itself, only imitating this based on a system-prompt and training.

This creates a problem for the epistemological status of the models as they are not inherently goal-oriented: It has no understanding of what it is doing other than minimizing its cost-function through trial-and-error learned action.
As Karl Popper famously noted, one cannot tell another to simply "observe!", as it needs to be directed towards something to make sense. For our model, the whole of its life-world will be created based only on the cost function and the initial input data. This is arguably different in humans, who, given their apperception (consciousness) and inherent goals of survival (see Darwin, Dawkins, etc.), will tend to have an automatic way of ordering their senses such that this makes sense, at least for human standards of intelligence.

The consequence of this lack of apperception and status problem is the faculties of judgement and reason cease to be relevant for a model. After all, storing particulars in concepts and reasoning about those implications requires some sort of abstraction where the abstraction is kept separate from the experience itself. Where humans would relate the particular to the concept and make a decision based on their knowledge of the concept, the Transformer model only has the particular. Given the way the model is subsequently not applying categories, but rather simply updating based on context, this equivalence between these dimensions (the status problem) leads to any decision being given in experience, rather than reasoned or judged as any action is derived from the immediate context rather than higher-order reason.

4.1.6: Practical Reason

We have now compared Kant's Transcendental System to the Transformer architecture and seen where the latter diverges from Kant's idea of human cognition. However, Kant posits at the end of the CPR that this cognition cannot be captured by the holistic Transcendental System alone, introducing Practical Reason as a missing link[40]. He argues that while the mind is limited to what is given in experience (Phenomena), it devotes considerable amounts of time to metaphysical claims about God, Freedom, and Immortality. This necessarily leads to contradictions as shown with his Antinomies and Paralogisms[41]. To Kant, striving to prove metaphysical speculation from an otherwise mostly rational mind is explained through the necessity of these beliefs for the Practical Reason[42], which is the mind applied to real-world decision making.

To explain the necessity of these beliefs, we start with Freedom and Morality. These are necessary, as without Freedom, reason wouldn't have anything to act on, and without Morality, reason wouldn't have any outcome to act towards. As such these two always co-occur and can be thought of as one Freedom/Morality duality, necessary for reason to be exercised and thus necessary for our human cognition.

---

[40] Kant, *Critique of Pure Reason*, A828
[41] Williams, Garrath. "Kant's Account of Reason." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2024 Edition. Sec 1.3.
https://plato.stanford.edu/archives/sum2024/entries/kant-reason/
[42] Kant, *Critique of Pure Reason*, A800

Kant then goes on to argue that a utilitarian-style optimizing often goes against our deontological duties as laid out by the Categorical imperative[43]. This means that a rational mind requires the belief that in the long term, keeping to those duties will lead to a positive outcome. This can then, in turn, only be explained by the belief a moral God would ultimately give you further future lives and the eventual happiness that you deserve based on your dutiful actions in the present[44]. Without these beliefs, the Morality/Freedom duality would not stand as no conflict would occur for Reason to extend to. This would in turn break down Practical Reason and with it the foundation Pure Reason has given us[45].

Importantly, Kant thinks these 3 beliefs in God, Freedom, and Immortality are both necessary to pre-suppose for Practical Reason -and thus human cognition- to function, *and* cannot be proven based on reason alone. As such, he claims both that Mathematics (as pure reason) cannot serve as a basis for Philosophy due to the impossibility of metaphysical knowledge[46], and that the aforementioned beliefs are necessary. From this comes his famous line from the second edition of the Critique "I had to deny *knowledge* in order to make room for *faith*"[47]. It is from this split that we may now see the implications of Kant's thoughts on our Transformer models.

Starting from God and Immortality before moving inward, it seems reasonable to assume that while a model may not have a Categorical Imperative to hold itself to, it can -and indeed must- be given another goal by its creators. This desired goal-state for the world to be in could then be compared to its current state to direct any action the model may take (i.e. to apply its Reason to). Where this will differ from human cognition is in the fact that there will be no conflict between its optimized utilitarianism and its deontological goals as they will be one and the same. This means that, similar in a human in Kant's view who wouldn't believe in God and thus not have this conflict, Freedom/Morality fall to the way-side, along with the ability to Reason.

This would be the end of the discussion for Kant: Given no conflicting desires, reason cannot be conferred, adding to yet another limitation that would have to be resolved for Transformer models to become capable of human cognition. Still, this remains a weakness in Kant's system as well. In the second Critique Kant realizes he cannot deduct freedom and has to appeal to it as an indubitable fact[48]. This means that effectively, we have just conferred a limitation on Transformer models which we cannot in truth deduct for humans either with the tools we have been using. As such, human Freedom and with it the credibility of these extra constraints are entirely outside of the scope of this thesis and may require detours towards debates on materialism, not to mention the fact that we could also capture such criticism by again referring to the lack of apperception as explained above. We can thus posit that for the purposes of our analysis, Kant is not sufficiently persuasive for us to expand our scope of limitations of Transformers, leaving only the restrictions as discussed prior in place.

---

[43] Kant, *Critique of Pure Reason*, A808
[44] Kant, *Critique of Pure Reason*, A813
[45] Kant, *Critique of Pure Reason*, A828
[46] Kant, *Critique of Pure Reason*, A727
[47] SEP, "Kant's Account of Reason," sec. 1.3.
[48] SEP, "Kant's Account of Reason," sec. 2.2

4.2: The Implications

After all of this we have an idea of the ways in which Transformer neural networks function differently from Kant's model on human cognition. We have set out a set of limits that pertain to specific states of development of the model and leniency given possible future expansion with infinite compute.. We can now see what kind of implications we have for the three fields we were interested in helping with this analysis, in Policy Making, Computer Science, and Philosophy (Epistemology).

4.2.1: The Implications for Policy Making

For policy making, the limits that we have discovered show that while the models may conceivably reach a level that could fool humans into thinking it performing reason (and may get as close as to functionally be able to pseudo-reason given the right goals and data), it will not pose any existential threat in itself as it is goal-less. Its incapacity to do reasoning limits the applications that it can have on novel discoveries outside of its training data as it cannot extrapolate with guaranteed logical rules at its foundation. This means that the calls for stopping the development of these models due to an existential threat seem to be out of line with the findings of this thesis. Rather, as the model comes closer to human-like intelligence when it comes to inference, productivity may be boosted by the assistance of such models, which may actually point towards a speeding up rather than slowing down of model development for the good of society.

This conclusion is based on the model in itself, however. When we expand the risks that these models could pose to include social risks around destabilizing society (by e.g. replacing jobs) these fears seem to be more grounded. Due to humans projecting consciousness onto chatbots, for example, negative social effects might follow in the form of excessive human-model relationships. The specifics of the social risks of these models are however outside of the scope of this thesis focused on the fundamental model structure, and I will thus not elaborate on them further.

4.2.2: The Implications for Computer Science

For computer science, especially the section on different model states seem to be of interest. After all, it is clear that as long as no underlying rules can be imbued into the system, like Kant's categories do, the models will be limited to their inferring, non-guaranteed functions. Due to this, model hallucinations, even when rare, could never be guaranteed to be eliminated due to the model's reasoning only being given in experience. While the model may thus become arbitrarily good at passing for a human, it will only ever be faking human-like intelligence.

Recently, Yann LeCun, one of the most important figures in AI-development and the Chief AI Scientist for Meta, admitted in an interview that he tells PhD students not to work on LLMs in their current form (including transformers). He argued that while incredibly useful for productivity, they are an off-ramp on the highway to true Artificial General Intelligence. Similarly, François Chollet, a senior in the Google AI team, echoed this sentiment, noting that he reckons OpenAI has set back the trajectory towards true AGI by 5-10 years by attracting a lot of the fields funding to a solution that will ultimately not get us closer. Elon Musk, however, is quoted as saying AGI could be achieved as soon as 2026. The findings of this thesis would suggest that LeCun and Chollet is correct, and that while useful, these models will not lead to anything truly resembling human intelligence.

4.2.3: The Implications for Philosophy

Lastly we have the implications for Philosophy, or more specifically, Epistemology. While this thesis was not necessarily aimed at providing insights for Epistemology, a rather interesting finding from the analysis follows when we look at the model in state three (infinite compute and data).

We have already shown that the model only functions on input and its manipulated subsequent states, so if this still allows the model to reach true reasoning capabilities given infinite compute and perfect data, this may point towards a transcendental type of reasoning not being necessary. After all, if the model could theoretically do reasoning and be Turing-complete without it, even if impossible in the real world, we could posit that perhaps the transcendental faculties used to reason are only truly necessary in practice rather than theory. This is significant as it may undermine a central point for Kant's system, as it is based on this necessity. Rather, we could argue that those faculties are used only as a practical alternative that "saves on compute" in comparison to the unrealizable experience-only option of our state 3 network.

5: The Conclusion

In this thesis we have substantiated and legitimized concerns over the rapidly improving state of the AI sector. We posited that while policy solutions existed for these concerns on technological advancement, these often require a constant, more-or-less fixed identity of the technology, which was lacking given this rapid improvement. As such, we argued that the best way to figure out a stable identity of AI systems would be to figure out their limits, regardless of compute which will improve further. To do this, we looked at the limitations of the technology, specifically Transformer systems, through the structure in which they operated, following in the footsteps of Kant's transcendental idealism. As such, we dove into both the architecture of Kant's system and transformer systems, to figure out where the differed and how transformers may (not) come to real knowledge. In this journey, we found that the specifics of the model state, i.e. how far we were willing to take our claim of compute being left aside, was a significant factor to the limits and concerns that could reasonably follow from the models. We then concluded that without infinite compute, i.e. in real-world scenarios, transformers would never be able to truly reach understanding due to a number of factors. Still, we found that simply that thought experiment had implications for not only Policy Making and Computer Science, but perhaps to the field of Epistemology itself, as we posited that experience-only architectures could theoretically achieve reasoning and thus knowledge-creation, potentially undermining the necessity of Kant's Transcendental architecture. Further research may focus on expanding this analysis to other model architectures, as well as exploring underlying assumptions that Kant has to make regarding Freedom.

Bibliography

Adejumo, Jaleel. "Gradient Descent From Scratch- Batch Gradient Descent, Stochastic Gradient Descent, and Mini-Batch Gradient Descent." *Medium*, April 16, 2023. https://medium.com/@jaleeladejumo/gradient-descent-from-scratch-batch-gradient-descent-stochastic-gradient-descent-and-mini-batch-def681187473

Ashish Vaswani, et al., "Attention Is All You Need."

Brook, Andrew, and Julian Wuerth. "Kant's View of the Mind and Consciousness of Self." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2023 Edition. https://plato.stanford.edu/archives/spr2023/entries/kant-mind/

Collingridge, David. "The Social Control of Technology," 1980. https://repository.library.georgetown.edu/handle/10822/792071

Copeland, B. Jack. "The Church-Turing Thesis." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2024 Edition. https://plato.stanford.edu/archives/spr2024/entries/church-turing/

Dahouda, M. K., and I. Joe. "A Deep-Learned Embedding Technique for Categorical Features Encoding." *IEEE Access* 9 (2021): 114381-114391. https://doi.org/10.1109/ACCESS.2021.3104357

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 16 (April 2018): E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Goel, Ashok K. "Looking Back, Looking Ahead: Symbolic Versus Connectionist AI." In *The AI Magazine/AI Magazine* 42, no. 4 (December 1, 2021): 83–85. https://doi.org/10.1609/aaai.12026

Hanna, Robert. "Kant's Theory of Judgment." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2022 Edition. https://plato.stanford.edu/archives/spr2022/entries/kant-judgment/

Kadtke, James B., and Linton Wells. *Policy challenges of accelerating technological change: Security policy and strategy implications of parallel scientific revolutions*. Center for Technology and National Security Policy, National Defense University, 2014.

Kant, Immanuel. *Critique of Pure Reason*. *Cambridge University Press eBooks*, 1998. https://doi.org/10.1017/cbo9780511804649

Kose, Parvez. "Feed Forward Neural Network — Explainable AI Visualization (Part 6)." *Medium*, February 20, 2023. https://medium.com/deepviz/explainable-ai-and-visual-interpretability-background-part-6-6467736f82b8

Kudina, Olya, and Peter-Paul Verbeek. "Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy." *Science, Technology, & Human Values* 44, no. 2 (2019): 291–314. https://www.jstor.org/stable/26637439

Lu, Yulong, and Jianfeng Lu. "A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions." Conference-proceeding. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, n.d. https://proceedings.neurips.cc/paper/2020/file/2000f6325dfc4fc3201fc45ed01c7a5d-Paper.pdf

Marr, Bernard. "A Short History of ChatGPT: How We Got to Where We Are Today." *Forbes*, February 20, 2024. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/

McLear, Colin. "Kant: Philosophy of Mind | Internet Encyclopedia of Philosophy," n.d. https://iep.utm.edu/kantmind/#SSH1ai

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." arXiv.org, January 16, 2013. https://arxiv.org/abs/1301.3781.

Raatikainen, Panu. "Gödel's Incompleteness Theorems." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2022 Edition. https://plato.stanford.edu/archives/spr2022/entries/goedel-incompleteness/

Shannon, Claude E. and Bell Telephone Laboratories, Inc. "Programming a Computer for Playing Chess." *Philosophical Magazine*. Vol. Vol. 41, March 1950. https://vision.unipv.it/IA1/ProgrammingaComputerforPlayingChess.pdf

Stang, Nicholas F. "Kant's Transcendental Idealism." *In The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2024 Edition. https://plato.stanford.edu/archives/spr2024/entries/kant-transcendental-idealism/

Stogin, John, Ankur Mali, and C. Lee Giles. "A Provably Stable Neural Network Turing Machine With Finite Precision and Time." *Information Sciences* 658 (February 1, 2024): 120034. https://doi.org/10.1016/j.ins.2023.120034

Vallance, Chris. "Elon Musk Among Experts Urging a Halt to AI Training," March 30, 2023. https://www.bbc.com/news/technology-65110030

Van De Poel, Ibo. "An Ethical Framework for Evaluating Experimental Technology." *Science and Engineering Ethics* 22, no. 3 (November 14, 2015): 667–86. https://doi.org/10.1007/s11948-015-9724-3

Van De Poel, Ibo. "Why New Technologies Should Be Conceived as Social Experiments." *Ethics, Policy & Environment* 16, no. 3 (October 1, 2013): 352–55. https://doi.org/10.1080/21550085.2013.844575

Varone, Marco. "The Best Part of Symbolic AI: Full Explainability." expert.ai, May 27, 2022. https://www.expert.ai/blog/the-best-part-of-symbolic-ai-full-explainability/

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." arXiv.org, June 12, 2017. https://arxiv.org/abs/1706.03762

Wang, Chi-Feng. "The Vanishing Gradient Problem - Towards Data Science." *Medium*, December 7, 2021. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

Watercutter, Angela. "Imagine if Joe Biden's AI Executive Order Were Inspired by 'The Terminator.'" *WIRED*, November 3, 2023. https://www.wired.com/story/imagine-if-joe-bidens-ai-executive-order-were-inspired-by-the-terminator/

Williams, Garrath. "Kant's Account of Reason." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2024 Edition. Sec 1.3. https://plato.stanford.edu/archives/sum2024/entries/kant-reason/

Ying, Xue. "An Overview of Overfitting and Its Solutions." *Journal of Physics. Conference Series* 1168 (February 1, 2019): 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Zhang, Jiawei. “Gradient Descent Based Optimization Algorithms for Deep Learning Models Training.” arXiv.org, March 11, 2019. https://arxiv.org/abs/1903.03614