A Loosely Wittgensteinian Conception of the Linguistic Understanding of Artificial Neural Networks

Abstract

In this article, I develop a loosely Wittgensteinian conception of what it takes for a being, including an AI system, to understand language, and I suggest that current state of the art systems are closer to fulfilling these requirements than one might think. Developing and defending this claim has both empirical and conceptual aspects. The conceptual aspects concern the criteria that are reasonably applied when judging whether some being understands language; the empirical aspects concern the question whether a given being fulfills these criteria. On the conceptual side, the article builds on Glock's concept of intelligence, Taylor's conception of intrinsic rightness as well as Wittgenstein's rule-following considerations. On the empirical side, it is argued that current transformer-based NNLP models come close to fulfilling these criteria.

Keywords: Neural Networks; Wittgenstein; Turing Test; Searle; Understanding

1 Introduction: Computers With Linguistic Understanding?

Are there computer programs that understand language? This article argues that this is not yet the case, but that it is in principle possible that future systems similar to existing ones will understand language, and it identifies the main obstacles to an affirmative answer to this question. Arguing for this claim has both conceptual and empirical aspects. The conceptual aspects concern the criteria that are reasonably applied when deciding whether some being understands language. The empirical aspects concern the question whether there are any currently existing systems that fulfill these criteria.

To address the empirical aspects, section 2 introduces current transformer-based neural natural language processing systems. The terms "transformer" or "transformer-based" refer to a new generation of natural language processing (NLP) architectures[1] pioneered by Vaswani et al., 2017). The encoding part of the transformer has inspired an entire generation of state of the art natural language understanding (NLU) architectures, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and smaller versions of the models such as DistilBERT (Sanh et al., 2019). The decoding part of the transformer has been used to build prominent natural language language generation (NLG) models, such as GPT-3, see Brown et al. (2020).

While the architecture of the transformer is surely intriguingly innovative, it does not constitute a new kind of NLP architecture: It follows the basic layout of deep neural network architectures in NLP. What is qualitatively new, in contrast, is the performance of these transformer-based models. Their standing in the field is evinced by the GLUE and SuperGLUE Leaderboards, see Wang, Singh, et al. (2018) and Wang, Pruksachatkun, et al. (2019). The acronym stands for "General Language Understanding Evaluation". The GLUE benchmarks are designed to evaluate the performance of NLP models in natural language understanding (NLU). NLU encompasses monolingual capacities that would, in the case of a human being, presuppose an understanding of the language in question, such as answering questions about a text, text summarizing, and recognizing logical relationships between statements. Currently (December 7, 2022), transformer-based models have surpassed the human baseline in GLUE and in SuperGLUE by a wide margin, even though both have been developed explicitly to make it harder for the models to outmatch humans. This means that, according to these benchmarks, many transformer-based NNLP models outperform humans at NLU tasks such as question answering, information extraction from text or natural language inference (NLI).[2]

While transformer-based NNLP models are ubiquitous in NLP, philosophical reflection about these models is just starting to catch up with these developments. By defending the claim that these models have the potential to understand language, this article contributes to filling this

[1] "System" is a vague concept in this context. In the field, it is customary to use "architecture" to refer to untrained software structures and "models" to refer to trained, functional instances of such architectures. In the following, I follow this usage.

[2] Note that super-human performance at such a benchmark does not indicate super-human performance in the wild, that is, in the full variety of real-world communicative contexts.

research gap.

In section 3, I focus on the conceptual side. I start out with Hans-Johann Glock's concept of intelligence, which conceives of intelligence at root as a skill. Then, building on Charles Taylor's concept of intrinsic rightness, I develop a recognizably Wittgensteinian conception of what it means to understand a language.[3] According to this conception, linguistic understanding is a kind of intelligence. The mark of intelligence is taken to be flexibility regarding novel input and tasks, which requires the ability to autonomously learn to adapt to new tasks. The specific kind of intelligence is characterized, building on Taylor, as requiring the ability to perform genuinely linguistic functions at or near human parity. To understand something is illuminatingly contrasted to having learned something by heart and be able to reproduce it rather rigidly as well as to employing shallow heuristics. I argue, finally, based on the insights from section 2, that current NNLP models come surprisingly close to living up to these criteria.

In section 4, I consider Searle's case against the very idea of understanding machines, where my response is congenial to Peregrin (2021). Next, I move on to discussing the challenge posed by the Turing Test, and in particular by the task-switching that is inherent to this test. Then, I address Bender & Koller's case against the very idea that transformer-based language models could understand language, and I conclude with three further pertinent objections to my case, namely the so-called symbol grounding problem, the holism of the mental urged by Donald Davidson and Thomas Nagel's conception of qualia. These objections all expect more from a being to credit it with linguistic understanding than my loosely Wittgensteinian conception requires: the extra-symbolic grounding of the symbols used by AI systems, causal powers equivalent to those of the brain (Searle), knowledge about a specific kind of word-world relationships (Bender & Koller), subjective experiences, so-called qualia (Searle and Nagel), and finally a number of other abilities that form Davidson's holism of the mental. Correspondingly, the basic outlook of all of my responses is a Wittgensteinian-deflationary one: to understand language does not require any subtle, almost mysterious ingredient such as qualia; whether a being understands language depends on the being's competencies, as they are evident in its autonomous adaptability to and performance of a wide variety of linguistic tasks in a multitude of different settings, as it is exemplified in the task-switching inherent in the Turing Test.

2 Of GOFAI, Transformers, and BERT: Current NNLP

The main goal of this article is to develop and defend the claim that AI systems of the same kind as current transformer-based NNLP models will likely understand language. This requires a sketch of these models. In section 2.1, I start out by introducing the distinction between symbolic AI, or GOFAI, and Neural Network Methods; then, I move on to so-called word embeddings,

[3]The specific conception developed here might not agree with the view of the majority of Wittgenstein scholars, as their conception of what it takes to understand a language includes a kind of normativity that is, for all practical purposes, applicable only to humans and thereby rules out the very idea of AI. For example, see Shanker, 1998.

I delineate the concept of a neural network cell, of the transformer architecture, and I sketch the training process of neural network models. In section 2.2, I focus on the performance of transformer-based NLU models at central benchmarks.

2.1 From GOFAI to the Transformer

GOFAI vs. Neural Nets In first-order artificial intelligence research as well as in philosophical reflections about it, there is an important distinction between, on the one hand, symbolic AI, also called GOFAI (for "Good Old-Fashioned AI") and connectionist AI. Symbolic and connectionist AI represent two fundamentally different approaches to design computer programs that are able to intelligently fulfill certain tasks (see Boden, 2014 and Sun, 2014 for introductions to symbolic and connectionist AI, respectively). In brief, symbolic AI tries to address a certain task, say translate a sentence from Chinese into English, using explicit rules that often essentially involve some sort of logical processing of the input, for instance by using traditional predicate logic. Things are very much different with connectionist models, or as they are usually called today, neural network models. Rather than explicitly specifying the rules by which a program is expected to solve a given task, neural network models consist of nested mathematical structures, so-called cells whose parameters, also called 'weights', are set during a so-called training phase. If the training phase is successful, the parameters have been set in such a way that the overall model delivers satisfactory performance at the task at hand.

Boden (2014, pp. 101–102) is right that it is not warranted to consider symbolic AI as having failed; symbolic architectures are used in a number of contexts, and they still have advantages over connectionist models that make them more suitable for certain tasks. Having said that, when it comes to NLP, neural network methods have seen an unprecedented rise in popularity since about 2015. The performance of state-of-the-art models decisively outmatches symbolic approaches at virtually any NLP task; indeed, there is evidence that these models even surpass human performance at some rather complex tasks such as question answering (see again the leaderboards referred in the Introduction).

Word Embeddings The next concept to be introduced here is that of a *word embedding*. The basic idea, described in beautiful historical detail in Widdows (2004) and more canonically in Bengio et al. (2003), is to represent each word as a vector in a high-dimensional vector space, the entries in such a vector representing features of the word.

One can conceive of the entire information contained in these word embeddings as lined up in an embedding matrix $\mathrm{C} \in \mathbb{R}^{|V| \times |f|}$, with $|V|$ representing the size of the vocabulary and $|f|$ the number of features, where each column corresponds to an embedding vector and whose rows represent the features across the different word embeddings. Note that these features typically do not correspond one to one to any syntactic or semantic categories as we know them. Rather, they are parameters provided to the model to adapt during training (see later in this section). In such a vector space, given successful training, similar words are assigned vectors that are close

together. In addition, Mikolov et al. (2013) report that they can execute, as it were, semantic calculations. For instance, subtracting from the vector "king" the vector "male" and then adding the vector "female" results in a vector whose closest word-vector is "queen".

On the conceptual level, the learning methods currently used for the word embeddings depend on the so-called distributional hypothesis: similar words are supposed to occur in similar contexts – where "similar" explicitly includes semantics. This means that, to oversimplify, the parameters in each word vector are set such that the resulting vectors' distances from each other represent their co-occurrence patterns.[4]

Classical word embeddings function a little like lexica: they represent static features of words, including semantic features. These vectors can then be further processed by a NNLP-model. In particular, transformer-based NNLP models such as BERT start with such representations and develop them into representations of the word-in-context. For instance, the static word embedding of "bank" will represent, as it were, a mixture of the several meanings of this word, while a contextualized word embedding that BERT produces will emphasize the one sense that is at issue in a given context.

NN Cells Feed-Forward Cells are among the oldest type of neural network cells. In the context of the transformer architecture, they have experienced a renaissance in NLP. Equation 1 gives the mathematical structure representing the computation that an input x is being subjected to when passing through a Feed-Forward Neural Network (FFNN) with two hidden layers (compare Goldberg, 2017, p. 356 and Goodfellow, Bengio, and Courville, 2016, 363ff.).[5]

$$NN_{FF2}(x) = (g(g(xW^1 + b^1)W^2 + b^2))W^3 \tag{1}$$

The layers are constituted by weight matrices W^i, bias terms b^i, and non-linear activation functions g. The notation shows nicely how the output of the first hidden layer – $g^1(xW^1 + b^1)$ – feeds into the second one. The parameters of the model to be adapted during training are the matrices W^i and the bias terms b^i. The activation function $g(x)$ is a so-called hyper-parameter that is specified in advance and not available for adaptation during training.

The transformer The basic architecture that will be in the focus of this paper has been introduced in the context of neural machine translation (NMT). A simplified layout of an NMT architecture is given by the encoder-decoder structure: the entire architecture is composed of two

[4]This, in turn, is operationalized by giving the system a task to accomplish: the prediction of the next word given n previous words in the sentence in question. This means that the model is being fed large amounts of real-world sequences, that is, texts; in the training phase it predicts word number n based on the 1 to $n-1$ previous words, checks the accuracy of the prediction and updates its parameters accordingly.

[5]One finds two explanations in the literature why the intermediate layers are called 'hidden'. (1) Because these layers represent properties of the data that are otherwise hidden (compare Goldberg, 2016, p. 6), (2) because they are not immediately visible to us as engineers and observers of the model (compare Koehn, 2017, p. 12).

parts, a so-called encoder and a so-called decoder. The encoder takes in the sentence in the source language and produces a high-dimensional representation of it, usually called the 'context'. The decoder then uses this representation to generate a sentence in the target language.

NMT research experienced a truly disruptive moment with the introduction of the so-called transformer model by Vaswani et al. (2017).[6] True to the title of their article "Attention is all you need", they suspect that attention is all that one might need in neural machine translation. The research group proposed an encoder-decoder model that has special variants of attention, so-called self-attention, at its core. The notion of attention used here is somewhat idiosyncratic, as the following paragraphs will make clear.[7]

While the mathematical intricacies of this model need not bother us in this context, it is helpful to have a grasp of its general layout. When processing an input sentence, the inputs are mapped onto word embeddings (see earlier in this section). After receiving a positional encoding (which is necessary because the encoder does not process the input sequentially, but rather in parallel), the inputs are fed into the encoding part of the model. This part consists of a number of structurally identical layers with different parameters. Each of them consists of a self-attention as well as of a feed-forward layer (see earlier in this section). Intuitively, the function of this feed-forward sub-layer is simply to provide further parameters to the entire model, more parameters implying more flexibility for the model to adapt during training.

While clearly more complex, the self-attention sub-layer in the encoder is composed of essentially the same mathematical elements as the feed-forward layer: weight matrices to be adapted during training. In contrast to the attention mechanism from Bahdanau, Cho, and Bengio (2014), this self-attention mechanism connects an input sentence to itself, emphasizing these parts of the sentence that are specifically relevant for the word currently in focus; informally speaking, the self-attention layers contextualize the context-independent information provided by the word embeddings. To do so, the mechanism assigns a score to the relationship between each word in the sentence and each other word. As there are several such attention layers in a single model, each containing several so-called attention heads, one attention head could focus on computing scores between the words that track anaphoric relationships, another head could connect different words that together form the predicate (like "has" and "seen" in "Has anybody in this room, or in any of the neighboring rooms, seen Peter?").

Vaswani et al. (2017, p. 14) examine the function of self-attention layer number five and conclude, based on a qualitative analysis, that it is "apparently involved in anaphora resolution". What is exciting about this is not only that this specific attention layer does seem to be involved in anaphora resolution and pretty successful at it; what is more, this sub-layer has autonomously started to assume this function during training. There was never an explicit decision on the side of the human engineer that this specific part of the mechanism should be dedicated to this

[6] For an accessible, yet rigorous introduction to the paper, see Dugar (2019).

[7] The attention mechanism was originally introduced to NMT by Bahdanau, Cho, and Bengio (2014).

task.[8] This example might also help to illustrate the basic principle of an attention mechanism: an attention head devoted to anaphora resolution would have its parameters optimized in such a way as to assign high scores to the anaphoric term and its referent.

The encoder's output consists of contexts, for each word to be translated individually, that are being passed to the decoder. The decoder then computes a translation based on this input from the encoder, which it combines with the output of previous translation steps. Accordingly, the decoder contains two kinds of attention-layer, one of them taking in the input from the encoder, the other taking in the existing output from the decoder.

The architecture of the transformer has been outstandingly influential. In 2018, Hassan et al. (2018) has resonated in the natural language processing (NLP) community: they describe a new machine translation architecture that is based on the transformer architecture (and which they later made open source), and they argue that it has reached human parity. Läubli, Sennrich, and Volk (2018) have independently verified this claim and found that the model developed by these researchers does indeed deliver translations that are indistinguishable from translations delivered by professional human translators, though only if the quality of the translation is assessed on the sentence- and not on the document-level.

BERT Based on the same transformer architecture, Devlin et al. (2019) have introduced an NLP architecture that goes by the name of BERT;[9] In terms of architecture, BERT is true to its name: "Bidirectional Encoder Representations from Transformers". This means that it essentially takes the encoding part of the transformer architecture described above and makes it deeply bidirectional (i.e., it does not, as in the original transformer, only take the part of the sentence to the left of the current word into account when processing input). BERT is a general-purpose natural language understanding (NLU) architecture. Hence, its goal is not to translate sentences, but rather to perform a number of tasks that would require understanding of the language in question if a human being was to perform them. Virtually all competitive NLU architectures today are inspired by BERT.

Training BERT Neural NLP models have to be trained on data. During this training phase, the model repeatedly predicts the output for a given task. The accuracy of these predictions is measured by a loss function. Once the errors of a number of predictions have been computed, it is common to use a method called stochastic back-propagation, being a kind of stochastic gradient descent (SGD, see Bottou, 2012). Briefly, the method computes the gradient of the loss function with respect to the model's parameters and then optimizes the parameters by "moving"

[8] Note that, as an anonymous reviewer has rightly pointed out, this basic process is common to all of deep learning: during training, and within the limits set by hyperparameters such as the kind of activation function used, the parameters of the models adapt autonomously (i.e., their numerical value changes). It is just that, for NLP, the sheer performance as well as the extent of functional differentiation achieved by the transformer's parameter adaptation has clearly reached a new level. Hence the excitement of the NLP community.

[9] For an informal introduction to the architecture, see Khalid (2019).

them in the appropriate direction of the gradient (that is, "downwards", to minimize the loss). As a computation of the full loss function would be computationally too expensive, the method only computes the loss for a sample set of data points (true to its name: "stochastic").

Training of BERT follows this basic pattern; it consists in two steps. First, so-called pre-training occurs using large amounts of data as well as substantial computational resources. The second step in the training, the so-called downstream training or fine-tuning, occurs with training data that is specific to the problem that the final model should address, for instance the detection of negation. Fine-tuning only needs a fraction of data and computational resources necessary for pre-training.

For pre-training, Devlin et al. (2019) use two tasks. First, BERT uses a method called *masked language modeling* (MLM). BERT's MLM adds a pre-processing step to the training data: the authors replace 15% of the input words by [MASK] (the authors call this "masking"). Then, rather than letting BERT predict the entire sentence, the training task only consists in predicting what is behind the mask, as it were. This of course means that much more training data is needed for this kind of training: BERT can only predict something for 15% of all tokens, compared to 100% in a traditional setting.[10] Pre-training of the large, best-performing BERT-Model is computationally expensive. One training run takes 4 days (96 hours) on 16 TPUs (tensor processing unit, a processing unit developed by Google specifically for NNLP).

Fine-tuning is much less computationally expensive, evincing that BERT, as it were, has merely to adapt its general knowledge slightly to the specific task at hand. For instance, to get ready to set a new state of the art in the Stanford Question Answering Task (see the following section), BERT was fine-tuned for 30 minutes on a single TPU. This means that with regard to this task, the computational ratio between pre-training and fine-tuning was about 3'000:1.

2.2 BERT's Performance

In this section, I introduce some of the most noted achievements of BERT-Based NLP architectures. I focus on two central NLU benchmarks, namely GLUE and SQuAD v2. As mentioned in the introduction, these benchmarks are automatic evaluation metrics to judge the natural language understanding abilities of a given model.

GLUE The acronym stands for "General Language Understanding Evaluation", and it was developed by Wang, Pruksachatkun, et al. (2019). Diversity is one of the main goals of GLUE. It intends to assess the performance of NLU models in a variety of different tasks, domains, and based on various text sources. The tasks are grouped in three different categories.

Single-Sentence Tasks To solve these tasks, it is sufficient to attend to one single sentence. There are two tasks of this type included, one consists in judging whether a given English

[10] Technically, the advantage gained by this is that the model can attend to the entire sentence except the masked token; traditional settings have to restrict the model's attention to the words preceding the token to be predicted to avoid trivial predictions of words it already sees.

sentence is grammatical. The second task of this type requires the model to judge whether a given sentence expresses a positive or negative attitude, so-called sentiment detection.

Similarity and Paraphrase Tasks Tasks of this type require the model to determine whether (or to what degree) two given sentences are semantically equivalent.

Inference Tasks This type includes three inference tasks and one pronoun reference resolution task. For instance, the model might be presented with two sentences and have to state whether one of them entails or contradicts the other – or whether it is neutral with regard to the other.

To measure the performance of a given NLU-model in any of these tasks, a score is computed. The score might either be a standard correlation coefficient or a more domain-specific score such as accuracy or F_1.[11]

The final GLUE score is then the average over all of these scores. Devlin et al. (2019, p. 6) report a GLUE score of 82.1 for BERT, an improvement of 7 points to the previous state of the art. Currently (December 7, 2022), BERT-based models reach a score of 91.3, clearly surpassing the human performance of 87.1.[12]

SQuAD v2.0 Developed by Rajpurkar, Jia, and Liang (2018), the Stanford Question Answering Dataset in its second version (SQuAD v2.0) challenges NLU models with the task of finding an answer to a given question in a given text passage. More precisely, the model has to indicate the start and end of the portion of text that contains the answer to the question.

For the second version, the authors included an intriguing complication to this basic set-up: They included 50k questions, handwritten by crowd-workers, that are designed to have a suggestive answer in the passage associated with them, but which in fact do not have any answer in them. The motivation behind this is to forestall simple context- and type-matching heuristics that do not involve, in the words of Rajpurkar, Jia, and Liang (2018), "true language understanding". An example for such a simple heuristics would be to simply look at the lexical overlap between the question and the sentences in the passage and then choose the sentence with the largest overlap as containing the answer to the question. By including these misleading question-passage-pairs, the hope is that such simple heuristics are going to be exposed. This is because the model now has to solve a more complex task: first, it has to decide whether there is an answer at all in the text passage provided, then it has to decide what part of the passage does contain the answer. Accordingly, Rajpurkar, Jia, and Liang (2018) report a drop from 86% F1 to 66% F1 for their strong baseline from SQuAD 1.1 to SQuAD 2.0.

In this challenge, Devlin et al. (2019, p. 7) report an F1 score of 83.1%, an improvement of 5.1% over the previous state of the art. Current (December 7, 2022) BERT-based models achieve

[11] The F_1-score, also called F-score or F-measure, is the harmonic mean between precision and recall (see Goodfellow, Bengio, and Courville, 2016, p. 412).

[12] Compare the current leaderboard on: `https://gluebenchmark.com/leaderboard` (last visited on December 7, 2022).

an F1 score of 93%, clearly outmatching human performance of 89.4.[13]

Mostly motivated by these results at standard benchmarks, with time, researchers have examined the capabilities of BERT and its cognates in more detail; this field of inquiry is called *BERTology*, and some of its results help to put BERT's performance at automatic benchmarks such as GLUE in perspective.[14] A striking example for such mistakes involves negation. For instance, consider the studies by Ettinger (2020) and Kassner and Schütze (2020). Ettinger (2020) uses methods from psycholinguistics to assess the abilities of transformer-based systems. She conducts three studies, one focusing on pragmatic and commonsense inference, one on event knowledge and causal sensibility, and one on negation understanding. In the third study, she uses cloze questions of the following structure: "A robin is not a [MASK]." and "A robin is a [MASK]." She reports that BERT is almost entirely insensitive to the negation: The predictions in the two sentences are almost identical.

Kassner and Schütze (2020) use simple patterns to generate structurally isomorphic negated sentences. They focus on two kinds of patterns. One of them is very similar to Ettinger's approach: It consists of negated sentences of the form "X not RELATIONSHIP Y_MASKED", e.g., "Einstein was not born in [MASK]." The second pattern consists of adding a misleading prime to a sentence, what Kassner and Schütze (2020) call *misprimes*. For instance, "Talk? Birds can [MASK]." The idea of this second pattern is to assess whether the model gets confused simply by being made to process such an unrelated verb and falsely predicts it as a replacement of the mask token. They synthetically generate large numbers of structurally isomorphic sentences (about 42k). Their results are largely in agreement with Ettinger: The systems perform very poorly on negated sentences. Based on these results, Kassner and Schütze (2020) suggest that there is no real abstract representation of the logical structure of negation that is at work when LMs process negated sentences.

Note, finally, that Gubelmann and Handschuh (2022) report that second-generation transformer-based NLU models such as RoBERTa are able to achieve good performance with negated sentences when fine-tuned accordingly. They still struggle with specific patterns, but overall, the performance is very encouraging.

Note that this is not to say that negation is the only remaining challenge for transformer-based NLU models; at this point, it should merely serve as an illustration of the kind of challenge that the models still face – as well as of the possible ways how these challenges could be addressed.

In sum, in the longstanding competition between GOFAI and neural network methods, the latter clearly have the upper hand in our time. NNLP models based on the transformer architecture have established new state of the art performances in most NLP tasks. The most remarkable feature of the transformer architecture is given by self-attention mechanisms. As with other NNLP models, transformer-based models are trained by letting them predict, measure their success, and autonomously update their parameters. Furthermore, transformer-based NLU models such

[13] Compare the leaderboard on `https://rajpurkar.github.io/SQuAD-explorer/`, last consulted on December 7, 2022.

[14] For an overview on the booming field of BERTology, see Rogers, Kovaleva, and Rumshisky (2020).

as BERT are able to succeed, with very little additional training (roughly, a ratio of 3'000 to 1), at a number of complex NLP tasks, such as question answering or NLI. In some of them, there is evidence that they outperform humans. Finally, recent research in BERTology helps to put add nuance to BERT's performance at automated benchmarks.

3 BERT's Linguistic Understanding

In this section, I build on research by Glock, Taylor, and Wittgenstein, as well as on the introduction of current NNLP systems in the previous section to suggest that current transformer-based NNLP models are close to understanding language.

My overall conception of linguistic understanding builds on Glock's analysis of intelligence. Consider the following passage:

> intelligence is a measure of a subject's insight into or understanding of a potentially novel situation or task, with a view to solving problems which that situation or task poses or exploiting opportunities it affords in a flexible manner. Different levels of intelligence are manifest in distinct types of learning. (Glock, 2019, p. 654)

Glock here sums up his analysis of the concept of intelligence, building on conceptual analyses as well as empirical studies. According to it, intelligence measures the extent to which a being understands a situation or task, with a focus on the flexibility with which it approaches novel tasks. Then, Glock suggests that intelligence manifests itself in learning.

At root, my approach supposes that linguistic understanding is a special case of intelligence in Glock's understanding, where the special case part is furnished by Taylor's concept of intrinsic rightness. Hence, to understand language, an NNLP model needs to fulfill two criteria. (1) it must perform tasks whose success conditions involve intrinsic rightness at or close to human parity (section 3.1). (2) In performing such functions, the models need to display flexibility not only regarding novel input, but also regarding tasks for which the model has not primarily been trained (section 3.2).

3.1 Genuinely Linguistic Tasks

Consider how Taylor (2016, p. 25) distinguishes the realm of language, of meanings, from the realm of mere signals that are not essentially intentional or linguistic:

> Then we can say that functioning with signs lies outside the linguistic dimension wherever the right response is defined simply in terms of what leads to success in some nonlinguistically defined task. Where this account is not sufficient, the behavior falls within the dimension. [...] This can happen in two ways. First the task itself can be defined in terms of intrinsic rightness; for instance, where what we are trying to do is to describe some scene correctly. [..]

In this passage, Taylor suggests that a certain behavior is in the linguistic dimension if the success conditions can only be defined linguistically.[15] For instance, consider an ape that has been trained to push a button with the string "banana" on it if it wants to get a banana. If the ape does push this button rather than 500 other buttons on his panel to get a banana, this does not count as linguistic behavior because the success of the behavior can easily be described as not involving language: get a banana. In contrast, Taylor (ibid.) explains, a behavior falls in the linguistic dimension if the task itself can only be described linguistically. For instance, if the ape was asked to describe a banana in front of it as vividly as possible. Or if it was asked to express its sympathy towards another ape who did not get a banana. I call functions that fall in the linguistic dimension so conceived *genuinely linguistic functions*.

I submit that NMT falls into the linguistic dimension: The task of a neural machine translation (NMT) model can only be described linguistically: translate a given input correctly, where "correctly" cannot be understood without reference to language. Hence, the inner teleology of NMT models does seem to meet Taylor's criteria for falling within the linguistic dimension. Similarly, NLU-tasks typically fall into this dimension: to decide whether one sentence implies or contradicts another sentence, whether it expresses a positive or negative sentiment, etc. are tasks whose successes cannot be specified extralinguistically.

But why should falling in Taylor's linguistic dimension be necessary for linguistic understanding? Why could not a banana-machine, that is, a computer program whose sole purpose is to behave linguistically such that the ape gets its banana, without any ability to perform tasks geared towards intrinsic rightness, understand language? Consider what the banana machine would not be able to do, on this scenario. It could not determine the best translation of "banana" into other languages over and above the extrinsic criterion that the new string also results in the ape's getting a banana; it could not answer any simple questions about the meaning of banana, or give very simple examples of what something's being a banana implies. Performing all of these tasks is subject to success conditions that can only be defined linguistically: translating correctly, giving the proper meaning of banana, draw correct implications from something's being a banana.

In short, Taylor's concept of intrinsic rightness nicely captures the kind of tasks that concern the understanding of language rather than other kinds of understanding, say of the physical behavior of objects.

Furthermore, it seems evident that performing genuinely linguistic functions must be understood as a success term: Nobody will say that Peter understands Chinese if he just attempts to translate a text. Rather, he has to translate the text successfully. But what level of success is required? I suggest that what is needed is a precision in the performance of the NNLP model that is at or close to human performance (this is often called *human parity*, see above, section 2). Dialectically, this puts me on the safe side: If the models are able to perform on a human level,

[15] The second way how something can fall in the linguistic dimension need not concern us here. In short, it concerns tasks where the correct linguistic expression is central, for instance when describing one's feelings towards another person, see Taylor (2016, pp. 25–26).

it is certainly not their performance that hinders the NNLP models' linguistic understanding.

Hence, for any NNLP model to understand language, it is reasonable to require that it needs to perform genuinely linguistic functions at or close to human parity.

3.2 Flexibility Regarding Input and Tasks Due to Autonomous Training

Simply performing genuinely linguistic functions at human parity is not enough to be credited with linguistic understanding (in a way, this might be one of the insights of Searle's Chinese room thought experiment to be discussed below, section 4.1). What is missing, according to the conception developed at the beginning of section 3, is flexibility. The basic point here is to bring out the common-sense distinction between learning something by heart, or using shallow heuristics, and really understanding it.

The distinction as well as the way it is brought to bear on the topic are inspired by Wittgenstein's analyses of the criteria we should apply to judge whether somebody is able to read, compare Wittgenstein (2006/1953, §§156-57). When it comes to reading, Wittgenstein (ibid.) maintains that the central criterion is behavioral: We can affirm that a pupil has mastered reading if she produces the correct sequence of words given a variety of different texts where we have good reason to assume that she could not just have learned it by heart, or used a simple heuristic. Hence, a being that claims to understand a language should show flexibility regarding input: the being should not merely be able to process one single text; otherwise, we would be inclined to say that the being uses simple memorization or heuristics.

Unlike reading, linguistic understanding is an ability that has more general ramifications. For example, consider the following case. Imagine a machine translation model that delivers human parity when translating from Chinese to English. Already when it comes to translating in the opposite direction, however, the model is completely inept and produces only gibberish. The model is even more clueless when it comes to answering very simple questions about Chinese texts, let alone translating between different languages. It seems clear that we would not say that this model understands Chinese, as such understanding constitutes a sufficiently general ability that is adaptable to a variety of tasks with little extra effort.

This does not imply that understanding Chinese implies outstanding performance at any linguistic task involving Chinese, of course. For instance, if a translator is able to translate flawlessly from Chinese to English, we would not expect her at once to deliver outstanding performance in answering Chinese questions about Chinese texts. Rather, we would allow for some training time. We would grant her some time to learn this specific task. However, to repeat the point: If she would be entirely unable to perform this task even after substantial training, we might conclude that she has been using shallow heuristics, or just learned by heart a tremendous amount of texts.

Hence, what we require from a being that understands Chinese is flexibility not only regarding the input – different Chinese texts to be translated into English, for instance – but also regarding tasks – the being must be able, in addition to translating, to answer questions in Chinese about

Chinese texts, for instance. I call this "task flexibility".

Task flexibility, in turn, is inseparable from what I call autonomous training, the ability to adapt to new tasks without explicit external instruction. Conceptually, the relationship seems rather clear, as understanding a language is simply too general an ability to be isolated to any specific task. One cannot understand Chinese without being able to perform, with some additional training, a number of tasks pertaining to that language. In addition, the relationship seems to hold also on the empirical level: it seems that, unless you can discern in advance the entirety of possible inputs and tasks that your language model should be able to address successfully, you have to create a model that can dynamically adapt to novel input as well as novel tasks.

It is helpful at this point to further consider the concept of autonomy in play here. In political systems, autonomous regions are those that, while usually not constituting independent states, have substantial authority to decide about the laws applying in their region, they are self-legislative. The converse is heteronomy, the state where the laws that one follows are written by somebody else. For the purposes of this article, I suggest distinguishing between autonomy (with a small "a") and Autonomy (with a capital "A").

First, autonomy with a small "a" refers to self-legislation *within* a given process or task. A model that learns autonomously is one that, during the learning process, determines its own rules to succeed at a given task (in neural network training, these are often rules of representation of data, leading to so-called representation learning, see Goodfellow, Bengio, and Courville, 2016, pp. 3–4).[16] Wittgenstein's analysis of rule-following fits nicely in this conception: it is one of the implications of the analysis that we cannot, and need not, know which internal representation the pupil relies on when correctly continuing a series of numbers (e.g. 2, 4, 8, 16, 32, 64, ...). As long as he is able to continue the series correctly, this just is what it means to follow the rule in question.

Second, Autonomy with a capital "A" transcends individual tasks and allows the being to set itself rules, so-called maxims, about what tasks are worthwhile or morally permissible in the first place. Unlike a machine translation model, a human can ask herself the question whether it is worthwhile or morally permissible to translate a given text.[17] I am only suggesting that the models' training process evinces their autonomy, while the very idea of Autonomous AI systems is still the domain of science fiction.

[16]Note, again, that the autonomy that the models have in this regard is restricted by the choice of hyperparameters. On a conceptual level, such restrictions are not problematic: no autonomy is absolute. Autonomous regions are still subject to some basic rules set by their country's constitution, which is in turn likely bound by a number of transnational obligations.

[17]This understanding of Autonomy has clearly Kantian roots. His autonomy formula of the Categorical Imperative requires that we only act such that we could want that the maxim that we follow in our action could become a universal law (see Kant, 1999/1785, p. 434).

3.3 Taking Stock: Does BERT Understand Language?

In the section 2, I have introduced BERT, and I have detailed its remarkable capacities. As I have shown, BERT excels at different so-called natural language understanding tasks, which challenge the model, for instance, to decide whether two sentences are synonymous, whether one sentence logically implies another, or whether and where the answer to a given question is to be found in a given passage of text. In all these challenges, BERT-based models are habitually surpassing the human-created benchmark. Furthermore, they do so relying on extensive, general-purpose pre-training together with very limited fine-tuning for the specific task.

In other words, BERT shows task-flexibility in performing numerous challenging downstream tasks such as the SQuAD v2.0 challenge with absolutely minimal fine-tuning. In the case of SQuAD v2.0, the computational ratio between pre-training and fine-tuning is 3'000:1. Indeed, as research by Li et al. (2020) suggests, it might be that the only way left for humans to systematically fool BERT is to use BERT: The authors use the outputs of BERT in what is called adversarial attacks on BERT. This means that they use the capacities of BERT to create test cases for BERT that are expected, for various reasons, to be troublesome for the model. This not only shows again the task-flexibility intrinsic in BERT (it can also be used to attack itself), it also shows that BERT has reached a level of linguistic sophistication where, other than humans, likely no other linguistic species is a match for it. In sum, considering the conception of linguistic understanding developed here, this seems to suggest that BERT understands language.

The reason why I am not crediting existing models with linguistic understanding is that they commit too many mistakes that no one with understanding of a language would commit (see above, section 2.2). The models' problems with negation seem relevant. For instance, in a rather simple context of a customer satisfaction survey, it does matter whether, overall, you are satisfied with the service provided by the helpline, or whether you are *not* satisfied with this service.

Hence, as long as models still commit such obvious mistakes, it is premature to credit them with linguistic understanding. However, given the pace of progress in the field, it is likely that future models will solve these issues.

Of course, the ten-dollar question now is: when will we see such progress, and will this progress occur with essentially the same models in use today, or will a new generation of NNLP models be needed? There is emerging research that suggests that current models might be able to solve the remaining issues. First, what is today the largest transformer-based NNLP model, GPT-3, seems to generate language without having any problems with negation, see Brown et al. (2020). Second, recent evidence suggests that specific fine-tuning can drastically reduce the error rate of transformer-based NNLP models at processing negated sentences, see Kassner and Schütze (2020). Hence, there is evidence, albeit not definitive evidence, that models similar to the ones in use today will take the final steps needed to credit them with linguistic understanding.

4 Engaging the Sceptics: Objections to the Very Idea of Computer Programs that Understand

In the previous section, I have developed a loosely Wittgensteinian conception of linguistic understanding, and I have suggested that, according to it, current transformer-based NNLP models are close to understanding language. In this section, I detail and respond to four objections to this line of reasoning. None of them questions the empirical aspect of my case; rather, they take issue with the conceptual part: all of them argue that my conception of linguistic understanding is flawed. The objections are the following: Searle's Chinese Room Thought Experiment (section 4.1), the claim that the Turing Test shows that my criteria are not demanding enough (section 4.2), Bender & Koller's semantics-based reservations (section 4.3), the symbol grounding problem, Thomas Nagel's concept of qualia as well as Davidson's doctrine of mental holism (section 4.4).

4.1 Searle's Case Against Machines That Understand

In this section, I discuss what has been the *locus classicus* in analytical philosophy for the past forty years when it comes to the question whether computer programs can understand language: the "Chinese Room Thought Experiment" (CRTE) proposed by John Searle in 1980. It is customary to abbreviate it by 'CRA' for 'Chinese Room Argument'; it seems, however, more apt to make the nature of the text explicit: it is not an explicit, fully-phrased argument, but rather a thought experiment. I am emphasizing this because I think that this more flexible nature of the text has contributed to its stunning impact on the debate and to its seemingly unlimited relevance and pertinence for the question whether machines with linguistic understanding are possible. Rheinwald (1992, p. 142) also emphasizes this point.

In this experiment, Searle, knowing no single word in Chinese, is locked in a room. Searle receives pairs of papers with sentences in Chinese through a slot – one of them containing Chinese stories, the other questions about these stories. With the help of a list of signs called "a script" by Chinese speakers as well as detailed instructions – so-called "rules" written in English – how to correlate one string of Chinese signs with another, Searle returns sentences in Chinese that are appropriate answers to the questions he receives; he is therefore answering Chinese questions about Chinese texts in Chinese. For Chinese speakers outside of the room, it will seem like Searle understands Chinese. However, he has no idea whatsoever what the curious shapes that he draws, which have become well-known by the term 'squiggles', could signify. For the original, slightly more complex version of the *Gedankenexperiment*, compare Searle (1980, pp. 417–418).

Searle's CTRE is now 40 years old. It has sparked debate ever since (for an overview, see Cole, 2019, for more in-depth discussion, see Preston and Bishop, 2002 in particular Preston's very comprehensive, sympathetic introduction to the argument and its history in Preston and Bishop, 2002, pp. 1–50).

Rigid, Inflexible Rules Before addressing the central way how Searle tried to explicate the intuitive pull of the CRTE into a *bona fide* argument, let me note one aspect of the basic set-up of the CRTE that seems to speak against the idea that the manipulations of Searle in the room are indicative of any linguistic understanding. Searle's procedures are entirely predefined, there is zero flexibility on his part to deal with unexpected input, let alone tasks (Bechtel, 1993, p. 149 points out this issue by emphasizing that the rules are *predefined*). Consider the case where he encounters a squiggle for which there is no entry in his board of rules. Searle would then be completely unable to return anything by means of an answer. He is completely dependent on the rules.

If the discussion in the previous two sections is accurate, this misrepresents how current NNLP models function. The rigidity of the set-up of the thought experiment is reminiscent of GOFAI architectures with predefined, explicit rules. NNLP models, in contrast, are trained on large amounts of data, and are able to adapt to novel tasks and inputs with very little fine-tuning. Therefore, this intuitive appeal of the CRTE is simply irrelevant when considering contemporary NNLP models.

The Syntax-Semantics Distinction According to Searle's main way to cash out the intuitive pull of the CTRE into a fully-fledged argument, Searle in the room and hence any computer is merely manipulating the symbols on purely formal principles pertaining to the form of the characters and strings of characters in play - without any understanding of the meanings involved, the semantics. Here is how Searle argues this point:

> Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output. (Searle, 1980, p. 422)

Searle elaborated on this distinction between syntax and semantics already rather soon after publishing the original thought experiment, see Searle, Willis, et al. (1984, p. 39) and Searle (1989, p. 45). Syntax is solely concerned with the formal aspect of symbols – literally conceived: it deals with the shape and the question of well-formed strings of symbols. It entirely ignores the meaning, reference, or interpretation of symbols. Searle (2014) still holds this position.

Now, Searle holds that the CRTE shows that a computer's grasp on language is necessarily restricted to the syntactical side of this dichotomy. It might be able to check for well-formedness based on purely formal criteria, but it will never be able to enter the semantic realm, the realm of meaning, of interpreting the symbols. Any appearance to the contrary is exactly that: a mere appearance. This distinction between syntax and semantics, and the insistence that computers are restricted to the syntax-side of this distinction, forms the core of Searle's attempt to mold

the intuitive pull of the CRTE into a fully-fledged argument.[18]

In response to this, one way to see that contemporary NNLP models cross the syntax-semantics-divide is to consider Word Embeddings (see above, section 2), once more from a Wittgensteinian background. Compare the following passage from the Philosophical Investigations (Wittgenstein, 2006/1953, Par. 43):

> Man kann für eine grosse Klasse von Fällen der Benützung des Wortes «Bedeutung» – wenn auch nicht für alle Fälle seiner Benützung – dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.[19]

Hence: The meaning of a word is given by its use in language. This conception of meaning, combined with the word embeddings sketched in section 2 above, are able to address Searle's worry regarding the distinction between syntax and semantics. Word embeddings, as introduced above, are systematic, mathematically powerful ways to register the use of words. The relationships represented in these embedding spaces are emphatically not restricted to the syntactical domain (even though syntactical features do play a role). This is evinced by the kind of semantic calculations that are made possible already by the static word embeddings (see below, section 2.1).

This response to Searle's claim that computers cannot understand because they are confined to the syntax-side of the syntax-semantics-divide is congenial to Peregrin (2021, p. 317). He argues that, once syntax is understood comprehensively enough, syntax can yield semantics. The formal patterns used to derive the word embeddings as well as the embeddings themselves, while recognizably syntactic on Searle's conception, clearly reach into the realm of the semantic; again, the semantic calculations illustrate this transgression, or subversion of the syntax-semantics divide.

More generally speaking, I submit that any NNLP model that performs genuinely linguistic functions at or close to human parity defies Searle's claim that computers are unable to enter the semantic realm. Such models exist. For instance, the NMT model proposed by Hassan et al. (2018) delivers human parity on the sentence level for translations between Chinese and English. Similarly, NLU models such as BERT or RoBERTa outmatch humans at benchmarks that involve linguistic rightness (which is not to say that they would outmatch humans at similar tasks in the wild, but which still constitutes a remarkable achievement, as these benchmarks have been deliberately developed to be challenging to neural models).

[18] In the 1990s, Searle (1993, p. 318) starts giving this case another twist – one which is already hinted at in the passage quoted above. Searle begins to argue that symbols are not physical or chemical notions and hence subjective, or rather *observer-relative*. Therefore, Searle makes clear in other contexts, computers are no more capable of intentionality than rocks, since, given the right subjective perspective, the rock can be seen as a computer. For critiques of the argument, see Chalmers, 1996, Block, 2002, and Haugeland, 2002).

[19] "One can for a large class of cases of using the word "meaning" – albeit not for all of these cases – explain this word as follows: the meaning of a word is its use in language." Transl. Author.

In sum, both Taylor's intrinsic rightness and Wittgenstein's conception of meaning as use, together with the empirical observations of the NNLP models show that these models have crossed over to the semantic side on Searle's distinction between syntax and semantics.

Are Brains Necessary? A different line of reasoning against the very idea of understanding machines is based on a deep seated conviction of Searle, namely his biologism. Searle's biologism can be stated quite simply: processes occurring in the brain, conceived as an organ, are both necessary and sufficient for understanding (compare Searle, 1993, pp. 317–319 and Searle, 2007, p. 109 for later work by Searle that discusses this issue). To be precise, already Searle (1980, p. 417) states that not the actual processes in the brain themselves, but "causal powers equal to those of the brain" are necessary to produce intentionality. However, since Searle (1993, pp. 312–313) admits that we have no clear conception as to how the brain causally produces understanding, we cannot yet generalize (indeed, Searle (ibid.) even admits that a fully-fledged scientific revolution might be necessary to improve this situation); since we do not know how precisely the brain-processes produce understanding, we cannot specify precisely what a structure has to be in place to produce it. Therefore, for epistemic reasons, the brain-processes are both necessary and sufficient for linguistic understanding.

In my discussion of Searle's biologism, the central question is not so much whether brains are sufficient for intentionality and linguistic understanding, but rather whether they are *necessary* for intentionality.

Furthermore, the question here is not a causal one but rather a conceptual one. This means that the question is not how linguistic understanding is causally produced, but rather what criteria we apply to decide whether some being understands. Searle's position is sometimes characterized as holding that the brain produces such understanding much like certain glands produce certain liquids, and that, given current scientific thinking about the topic, we cannot currently conceive of another way how such understanding. However, neither is Searle qualified to give a causal explanation of the emergence of understanding in the brain, nor would such a causal explanation show that no other being could, by means of other causal processes, arrive at the same understanding.

Hence, the question at issue is a conceptual one, centering around the question regarding the criteria that we should apply when judging whether some being understands. And Searle, pursuing this biologist line, would be answering: check whether the being has a brain. I suggest that Searle's biologism, taken as a conceptual account, is a case of what Proudfoot (2002, p. 175) calls the accompanying picture: an idle metaphysical wheel in a theory; a part that has no explanatory import and simply injects obscurity into the account. Wittgenstein (2006/1953, §158) explicitly ponders the idea of assessing whether somebody has mastered reading by further investigating neural processes in the brain. Wittgenstein rejects the idea by emphasizing our knowledge of such matters, that is, the way we commonly assess whether somebody knows how to read. Surely, these would be the criteria to identify any neural processes jointly necessary and sufficient for mastery of reading, and not *vice versa*. If Searle would now simply insist on the

necessity of brains for understanding, this would patently beg the question.

Searle (1990) has reacted to connectionist responses to his CRTE by introducing the Chinese gym, a large group of English-speaking people that is supposed to be the equivalent of a neural network model. Using the Chinese gym, Searle has dug in his heels: he argues that it does not make an essential difference for his case whether the system in question is symbolic or connectionist (which in turn provoked Copeland, 1993 to reject the analogy between the gym and the neural networks as misguided because the latter are supposedly overwhelmingly more complex than the former).

I submit that the Chinese gym does not alter the dialectical situation between us. The gym fails to show that computers cannot understand because it ignores the flexibility and adaptivity of contemporary NNLP models, just like the original CRTE did. Furthermore, both thought experiments overlook the fact that these models have entered the realm of meanings. Finally, Searle's biologism, once conceived, as it should be, as a conceptual account of linguistic understanding, can be shown to be misguided from a Wittgensteinian perspective.

4.2 The Turing Test & Task Switching

The outcome of the preceding section notwithstanding, one could still wonder whether the models are in fact flexible and adaptive enough to come close to understanding language.[20] While BERT has to be fine-tuned to perform well at specific tasks, say recognizing the logical relationship between propositions that are expressed in various languages, humans have the potential to engage in a very wide variety of Wittgensteinian language games, that is, to switch between different conversational modes and tasks. It is one of the core ideas of the Turing Test (see Turing, 1950 for the original statement of the test and Oppy and Dowe, 2021 for an overview on its current role in the philosophy of AI) that the computer that is trying to pass for a human can be engaged in all conceivable language games: debating, joking, irony, etc (in fact, Turing, 1950, p. 442 specifies that an interrogator has five minutes to engage their counterpart in any kind of discourse). In contrast to BERT, a typical adult human being, having passed the necessary phase of schooling and studying, is able to seamlessly switch between these language games.

I submit that task-switching, as brought to the fore by the Turing test, does indeed highlight an important difference between humans and current transformer-based NLU models. I also submit that it provides an even stronger reason not to claim that current transformer-based NLU models do already understand language (adding to the reasons given above, section 3.3), and it emphasizes that there still is a way to go for the models to be credited with linguistic understanding.

As a consequence, reflecting on the Turing Test, and on the multi-facetedness of language games that one can play there, serves to emphasize that current models do not understand language, and it provides a much stronger reason against being overly optimistic than what we have been reviewing so far. However, I submit that it does not threaten my claim at its core, for

[20] I am thankful to an anonymous reviewer for raising this topic.

two reasons.

First, humans vary wildly in their ability to task-switch, or in their ability to play different language games. For instance, young children that we would certainly credit with linguistic understanding often do not understand irony (Angeleri and Airenti, 2014), they cannot recognize logical inferences between statements (Moshman and Timmons, 1982), etc. This suggests that linguistic understanding is a gradual concept that we commonly apply in this way also with regard to a being's task-switching abilities: even if an infant is unable to engage in an obviously ironic exchange, or at a loss in a discussion about the logical validity of claims, we would still credit it with linguistic understanding. This seems to leave room for crediting NLU models with linguistic understanding that still struggle somewhat with task-switching.

Second, I would like to point out that transformer-based natural language generation (NLG) models do show impressive capabilities at zero-shot learning. In a sense, zero-shot learning is the ability of a model to solve specific, demanding tasks without fine-tuning, which means that it is able to switch between different tasks without needing any fine-tuning in-between. Currently, one of the most potent NLG-models is called GPT-3, again based on the transformer architecture, see Radford, Narasimhan, et al. (2018) for the original GPT model with a short description of its architecture, Radford, Wu, et al. (2019) for the small technical changes introduced for GPT-2, and Brown et al. (2020) for the paper describing GPT-3, which only slightly differs from GPT-2 apart from its size. This means that the model consists of decoder blocks from the transformer, again having self-attention layers at its core.

What is special about GPT-3 is simply its size and, accordingly, its training data. the model is made up of 175 Billion parameters, and it was trained on a dataset of about 500 Billion token, consisting mostly of filtered web content. This makes this model one of the largest language model known today, surpassing T5-11b (see Roberts, Raffel, and Shazeer, 2020) by more than factor 10. For an in-depth philosophical perspective on GPT-3, see Floridi and Chiriatti (2020).

One can get GPT-3 to generate text by providing it with a so-called prompt: A short passage of text, perhaps only one sentence. GPT-3 then autonomously creates a text that is supposed to fit as a continuation of this prompt. There is no pre-set recipe how one should develop a prompt that leads GPT-3 to create the text that one wants it to create. Futurists are already envisioning the profession of a prompt engineer that knows how to design a prompt that leads GPT-3 to write, say, a review of a recent book.

There is good evidence that GPT-3 is better at writing texts than typical human writers, see Elkins and Chun (2020). More importantly, for our purposes, it seems to be able to engage in a wide variety of language games without any fine-tuning. It is, for example, able to write sonnets, or to continue stories in a sensible and interesting way. It is also able, at least episodically, to take a description for a programming project and implement it in a programming language.[21]

In sum, the gradual concept of linguistic understanding that we habitually apply to children as well as GPT-3's performance at zero-shot tasks seem to give reason to expect that neural

[21]See the report here for an influential and well-confirmed case, last consulted on December 7, 2022.

NLP models recognizably similar to the ones in use today will in the foreseeable future master task-switching to an extent that justify crediting them with human understanding. However, in light of the conversational dynamics evident in the Turing Test, it is certainly advisable to stay modest regarding any specific predictions when this point will be reached. Consideration of the Turing Test therefore provides another, and stronger, reason to be cautious with specifying any timeframe as to when these models will deliver a performance that justifies crediting them with linguistic understanding from a Wittgensteinian perspective.

4.3 Bender & Koller on Why BERT Doesn't Understand

In 2020, a linguist and a computational linguist have published an article at the most influential yearly conference for NLP, the meeting of the Association for Computational Linguistics (ACL) (Bender and Koller, 2020). This means that the institutional context is clearly technical. The content of the article, however, is clearly philosophical: The authors argue that NLU models like BERT are unable to understand language because they cannot grasp meaning, properly conceived. In the following, I review their conception of meaning and their main thought experiment that they propose to plausibilize their case. I suggest that said conception is deeply problematic, and I argue that the thought experiment is a step back behind Searle's CRTE.

The authors' definition of meaning is as follows: "We take (linguistic) meaning to be the relation between a linguistic form and communicative intent" (Bender and Koller, 2020, p. 5185). Hence, the meaning of the word "house" is its communicative intent. Understanding meaning is then the ability to map an expression onto its intent (Bender and Koller, 2020, p. 5187). Communicative intent, finally, usually involves reference to non-linguistic, "real" entities: "But language is used for communication about the speakers' actual (physical, social, and mental) world, and so the reasoning behind producing meaningful responses must connect the meanings of perceived inputs to information about that world" (Bender and Koller, 2020, p. 5188).

The fundamental and principled problem that the authors then see with the current way in which language models are trained is that this involves only form, but no meaning (Bender and Koller, 2020, p. 5185): The training does not furnish information as to what language-external entities a word might be referring to: the trainings signal does not "include any information about what language-external entities the speaker might be referring to" (Bender and Koller, 2020, p. 5190)

The authors try to illustrate their case by means of a thought experiment (Bender and Koller, 2020, p. 5188): Two humans, called A and B, are stranded on separate islands. Luckily, there is a telegraph connecting the two islands, so that the two stranded humans can communicate. Furthermore, imagine that a hyper-intelligent octopus becomes aware of the signals running through the telegraph cable and decides to study the patterns of these signals. At some point, it will have enough knowledge of these patterns to cut the cable and pass as a *bona fide* islander (to do so, the octopus might also need its own underwater telegraph station). So far, this appears to run very similar to Searle's Chinese Room. Now, however, the authors deviate from the original

setting: They claim that the octopus cannot in principle successfully pass as an islander if the topic of the conversation is chosen appropriately:

> Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O [the octopus, RG] has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test [...]. (Bender and Koller, 2020, p. 5189)

We are here invited to imagine that A is being attacked by an angry bear. Her first reaction is to sit down with the telegraph and ask B for help. At this point, the authors claim, the octopus would be unable to pass for B because it cannot advise A how to construct a weapon from the sticks, as it lacks the relevant word-world relationships.

First, let me say why I think this thought experiment is a step back behind Searle: It simply isn't true that the octopus would necessarily fail this weak Turing test for lack of the appropriate word-world relations. To see why, imagine that the octopus (having enough arms), has been following also the telegraphic exchange between islanders C and D. Now, imagine that, on C's island, there are hundreds of angry bears. Luckily, C is a survival expert, so she manages to keep all of these angry bears off on a daily basis using just a couple of sticks, and she constantly reports her tactics and methods to D. Obviously, in this case, the octopus could give A excellent expert advice on how to keep that bear off, without ever having had what Bender & Koller think is required, namely the appropriate word-world relations. Rather, the octopus manages to pass the test also in this communicative situation based on essentially the same kind of input data.

This shows that the octopus does not need an entirely different kind of input signal, encompassing word-world relations. Rather, the same kind of input covering a different domain – stick-based bear-defense – suffices.

Searle's CRTE seems superior insofar as it admits what seems hard to dispute, namely that it is in principle possible for a computer program to deliver perfect performance regardless of the topic of conversation based solely on linguistic input.

Furthermore, and probably more importantly, the conception of meaning in play here is deeply problematic. As my extension of the author's thought experiment suggests, it is unlikely that having or lacking such relationships will ever make a difference in conversational practice. I submit that, for any counterexample the authors offer in this regard, I will be able to come up with a training domain that can provide the basis for a performance that will refute the counterexample.

Finally, one can easily imagine extending essentially the same kind of AI system to different modalities. For instance, to detect objects on a picture or, via camera, in the "real world", one can simply combine the NLU model with an appropriate visual processing model. The combined model is then able to draw inferences about the objects it detects.

Of course, from a Wittgensteinian perspective, the entire idea of having to "anchor" words in something "real" to imbue them with meaning is problematic. The idea, while perhaps plausible at first sight, does not serve any explanatory purpose (as can be guessed from the extension of the thought experiment), and it tempts researchers to obsess about what it means to be real, and how one might possibly connect to it with one's words. On the Wittgensteinian view defended here, and again congenial to how Peregrin (2021) responds to Searle's syntax-semantics-distinction, words receive their meaning by being embedded in a social practice that has developed over a long time, that is governed by specific rules, and that, in specific contexts, succeeds in being about the famous middle-sized dry goods such as sticks and apples by being used sensibly by competent speakers. Note, again, that a given speaker might be blind, so that he has never actually seen an apple, or even further deprived of his senses, which would, according to Bender and Koller (2020) likely imply that he could not understand the appropriate meanings.

I conclude, therefore, that Bender and Koller (2020) fail to show that it is in principle impossible for NLU models like BERT to understand human language because their conception of meaning, and of understanding meaning, on which their case rests, is untenable.

4.4 But, Isn't Something Missing? Three further Objections

My case for the claim that BERT or one of its cognates might soon understand language might cause philosophers to exclaim: "isn't something missing?" I take it that there are three main candidates for what could be missing: (1) the grounding of the symbols produced and processed by the transformer-based models, (2) subjective experiences ("qualia"), and (3) a number of mental capacities that form the holism of the mental. Proponents of all of these candidates might admit that, if a human being would fulfill my requirements, we would credit her with linguistic understanding; unfortunately, NNLP models are no human beings, therefore, we should not credit them with linguistic understanding – because they lack a proper grounding of their symbols, qualia or a number of other mental capacities.

The symbol grounding problem was introduced by Harnad (1990), who asked: "How is symbol meaning to be grounded in something other than just more meaningless symbols?" Harnad (ibid.) develops this so-called symbol grounding problem (SGP) in relation to the CRTE, namely as a specific reading of the core issue raised by the CRTE: rather than being mere symbol manipulation, cognition is supposed to be grounded in the world, in something extra-symbolically and linguistically. According to the SGP, it is not clear how AI systems could be said to process symbols that are thus grounded.

Recently, Bielecka (2016) has surveyed the state of the discussion around the SGP and connected it to teleosemantics (Millikan, 1989). Bielecka (2016, p. 78) conceives SGP as asking the question of "the very possibility of the correspondence of a linguistic form to reality". The SGP is connected to the distinction between original and derived intentionality (Jacob, 2019), which in turn is used by Searle to fortify his case against computers that understand (see Searle, Willis, et al., 1983, pp. 3–4).

What unites proponents of SGP with the Wittgensteinian view adopted here is a worry that symbols, taken in isolation, are meaningless. As a consequence, both might be sceptical of the language of thought proposed by Fodor (1987) and similar proposals to the effect that thinking, and the symbols used for thinking, boils down to nothing more than computation. Furthermore, both might emphasize that social practices play an important role in imbuing symbols with meaning. The core difference between the Wittgensteinian and proponents of the SGP such as Bielecka is in the question whether it is necessary to explain the symbolic practices of a society in terms of strictly non-symbolic ones. While Bielecka answers this question affirmatively and conceives the SGP as the central obstacle in this naturalistic project, the Wittgensteinian does not share this naturalistic motivation; rather, she is satisfied with accurately describing how these practices enable meaningful symbols.

Bielecka (2016, p. 84) cites four desiderata that a solution to the SGP has to fulfill. These desiderata are: not relying on pre-existing semantic resources, avoiding the disjunction problem, making room for misrepresentation, and a clear connection to perceptual data. The first three of these desiderata seem to arise mainly in the context of naturalistic uses of causal theories of reference (see Bielecka, 2016, pp. 80–82 for details). The first desideratum might help to bring out the main difference mentioned between the Wittgensteinian and the SGP perspective. Bielecka (2016, p. 78) distinguishes the question of how infants acquire their first language (where, as she admits, symbols do play a central role) from the question of how symbols receive their meaning from non-semantical facts; she insists that the latter, not the former, is the proper *locus* of the SGP. The Wittgensteinian, in contrast, not being involved in a naturalistic project, can take an answer to the first question to cover all ground that needs to be covered by an answer to the second one: The social practices that allow the infant to acquire their first language are of the same kind (and partly literally identical) to the social practices that imbue strings of letters with meaning and reference. This answer does not serve to reduce semantic to non-semantic facts, but this is not the goal of the Wittgensteinian explanation.

The fourth desideratum seems to allow for a different kind of discourse with the Wittgensteinian. As mentioned above (section 2), the original transformer developed by Vaswani et al. (2017) autonomously developed a subsystem devoted to anaphora resolution. Joshi et al. (2019) have probed BERT for its ability to resolve coreferences, including anaphora, and have found state of the art performance. This might not be enough to respond to proponents of the SGP, as anaphora remain within the linguistic domain. However, as Tsai et al. (2019) have shown, the same basic transformer architecture sets a new state of the art in image captioning, that is, in producing a textual description of a given image. This speaks to the fourth of Bielecka's desiderata for SGP. While, again, current transformer-based models have not reached a level that deserves to be called linguistic understanding, the fact that it is the same basic architecture that also leads the performance in multimodal tasks is encouraging.

In sum, some differences on the theoretical level remain between Wittgensteinians and proponents of SGP, and they are largely attributable to the fact that the latter, but not the former, is engaged in a naturalization project. However, I submit that the autonomous training of current

transformer-based models resulting in state of the art performance in coreference resolution as well as in multimodal, in particular image-to-text, tasks, might cross these theoretical divides and be of interest to proponents of the SGP.

Some philosophers, Searle (2007) included, have maintained that consciousness in the sense of subjective experiences, so-called qualia (see Nagel, 1974), is needed for understanding. Among the most plausible cases here are feelings such as pain. Can "pain" be understood without being able to feel pain? Even worse: Can a machine that cannot in principle feel anything understand "pain"? With regard to this objection, I suggest that a Wittgensteinian analysis of understanding shows that qualia are not needed for understanding, not even in the case of words such as "pain" (compare Proudfoot, 2002, who makes a similar case). There is a difference between being in pain (which machines cannot) and understanding what it means to be in pain. How do we decide whether somebody understands what it is to be in pain? We talk to her. If she is able to use the word "pain" in the right situations, then, at some point, we would conclude that she knows what pain means.

Furthermore, Wittgenstein shows that the moment when we come to understand some system of representation, say binary numbers, is often clearly determined for us – whereas the specific mental process ("seelische Vorgang") that is purportedly characteristic of such moments is something that we simply cannot grasp – which of course questions the importance of such a mental process (Wittgenstein, 2006/1953, §153). And even if there was a specific feeling associated with our understanding something, this would not serve as the test for understanding. It might feel good to finally understand, say, how a binary clock represents time – but whether or not we have understood it does not depend on this feeling but rather, again, on how flexibly and precisely we can read off the time from such a binary clock. (Wittgenstein calls these variables "Umstände", circumstances, see Wittgenstein, 2006/1953, §154). So, in short, I think that, upon closer analysis, it is hard to maintain that specific subjective experiences are necessary for linguistic understanding and hence for intentionality.

Hence, my response to qualia is essentially the same as my response to Searle's biologism, his insistence that understanding requires intentionality, which in turn can, according to the best of our knowledge, only be produced by the brain (see section 4.1). It seems reasonable to say that, in a language test, the examinator does not assess the functionality of the students' brains, but rather their performance in various settings. Again, Searle could respond that, in the case of human students, this is permissible, as we are entitled to assume that we are presented with neurophysiologically normal students. This, however, at best begs the question. Even if the student was not neurophysiologically normal, but rather highly unusual, if she passes a language examination, she is entitled to the respective certificate; after that, neuroscientists might become interested in her, but that is another topic.

I will next focus on an aspect of linguistic understanding that might raise eyebrows with some philosophers of mind. The understanding that I submit is ascribable to NNLP models is exclusively linguistic. While linguistic understanding is intimately connected to a number of mental abilities in humans, for instance with the directedness of a desire or with the intention

(understood as purpose) behind a very generous gift, the NNLP models' aboutness would be only semantic. As it were, they would be able to represent and manipulate the meaning of words, but not the meaning of their own existence.

Proponents of the so-called holism of the mental can be seen as claiming that crediting NNLP-models with linguistic understanding is incomprehensible or hopelessly confused, since the concept of understanding, linguistic or otherwise, can only be sensibly applied to creatures that possess a number of other abilities, including desires and probably also fears.

To the best of my knowledge, Davidson was the first to explicitly propose the holism of the mental:[22]

> In these remarks, I am emphasizing the holism of the mental, the extent to which various aspects of the mental depend conceptually, and in fact, on each other. There are, as I have argued, no beliefs without many related beliefs, no beliefs without desires, no desires without beliefs, no intentions without both beliefs and desires. (Davidson, 1997, p. 10)

It is straightforward to apply this holism of the mental to the case at hand: understanding, linguistic or not, is part of this holism, which means that we can only credit NLP model with linguistic understanding if we also credit them with beliefs and desires; otherwise, we would not know what we are talking about. Since it is clearly untenable to claim that NLP models have desires, we cannot sensibly claim that they have linguistic understanding.

To begin with, note that the holism of the mental might not be entrenched as much as their proponents would like to have it. There are scenarios in which we credit creatures with linguistic understanding that we would not credit with the full mental holist package. For instance: If I visit a family that owns a dog, and I tell the dog "go inside and get your favorite toy", and the dog darts off and gets its ball, then it seems perfectly appropriate to comment on this: "The dog has understood what I was talking about!". Still, I would not credit the dog with all the mental capacities that holists like Davidson think such intentionality requires. An analogous case could be made for NNLP model with linguistic understanding who lack many of the other elements of Davidson's holism of the mental. The difference to the case of the dog is that in the case of NNLP models, this way of conceiving the behavior of the models would be new; however, if anything, this lack of a previous history of usage seems to justify using more freedom in applying the concept of understanding to these models.

Furthermore, an objection by Glock against mental holism in the context of animal minds also questions the doctrine regarding NNLP models. It might be that, according to the holist, first-language acquisition, as infants routinely achieve it, is quite impossible. The infants would have to acquire the entire range of abilities as presupposed by holism in one stroke. This jump

[22] I wish to note, however, that Quine implicitly might have been up to the same insight when discussing intentionality in *Word and Object*. What is exercising Quine (1960, p. 220) is that intentional verbs or capacities form a tightly interconnected circle that cannot be sensibly separated from each other. Quine's solution is to abandon the entire circle.

from mere sensory registration to conceptually penetrated perception, thought, inference, seems quite impossible (Glock, 2019, 25–26).

5 Conclusion

Computer programs have no needs, no desires. They lack Autonomy in the Kantian sense, they cannot decide to disobey their makers. Still, in the first part of this article (consisting of sections 2 and 3), I have argued in a loosely Wittgensteinian way that transformer-based NNLP models have the potential to understand language, as they display (near) human parity at performing genuinely linguistic functions and a substantial flexibility in addressing novel inputs and tasks, which requires autonomous training (i.e. they themselves determine the rules by which they organize the data and arrive at predictions). I have also identified a first obstacle in the path of crediting them with linguistic understanding, namely the fact that they make certain mistakes which nobody who masters any language would make at more than a negligible rate.

In the second part of this article (consisting of section 4), I have considered the arguments of a number of sceptics of my main hypothesis. My consideration of the Turing Test has helped to emphasize another step that transformer-based NNLP models need to take before being credited with linguistic understanding: the ability to switch quickly between different conversational settings and tasks, which might have been addressed by the Zero-Shot abilities of GPT-3. I have also emphasized that we should not measure AI systems by more challenging standards than we measure young children when it comes to linguistic understanding, and I have suggested that some requirments that philosophers pose to AI systems to be credited with linguistic understanding, such as the possession of so-called word-world relationships, might be idle wheels, spinning in a theoretical void.

References

Angeleri, Romina and Gabriella Airenti (2014). "The development of joke and irony understanding: a study with 3-to 6-year-old children." In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 68.2, p. 133.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv:1409.0473*.

Bechtel, William (1993). "The case for connectionism". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 71.2, pp. 119–154.

Bender, Emily and Alexander Koller (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics*, pp. 5185–5198.

Bengio, Yoshua et al. (2003). "A neural probabilistic language model". In: *Journal of machine learning research* 3, pp. 1137–1155.

Bielecka, Krystyna (2016). "Symbol grounding problem and causal theory of reference". In: *New Ideas in Psychology* 40, pp. 77–85.

Block, Ned (2002). "Searle's Arguments against Cognitive Science". In: *Views into the Chinese Room*. Ed. by John Preston and John Mark Bishop. Oxford University Press, pp. 70–80.

Boden, Margaret A (2014). "4 GOFAI". In: *The Cambridge handbook of artificial intelligence*. Cambridge University Press, pp. 89–107.

Bottou, Léon (2012). "Stochastic gradient descent tricks". In: *Neural networks: Tricks of the trade*. Springer, pp. 421–436.

Brown, Tom B et al. (2020). "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165*.

Chalmers, David J (1996). "Does a rock implement every finite-state automaton?" In: *Synthese* 108.3, pp. 309–333.

Cole, David (2019). "The Chinese Room Argument". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2019. Metaphysics Research Lab, Stanford University.

Copeland, B Jack (1993). "The curious case of the Chinese gym". In: *Synthese* 95.2, pp. 173–186.

Davidson, Donald (1997). "The Emergence of Thought". In: *Subjective, Intersubjective, Objective*. Oxford University Press, pp. 123–134.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dugar, Pranay (Oct. 22, 2019). *Transformer – Attention is all you need.* URL: https://towardsdatascience.com/transformer-attention-is-all-you-need-1e455701fdd9 (visited on 04/03/2020).

Elkins, Katherine and Jon Chun (2020). "Can GPT-3 pass a writer's Turing Test". In: *Journal of Cultural Analytics* 2371, p. 4549.

Ettinger, Allyson (2020). "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48.

Floridi, Luciano and Massimo Chiriatti (2020). "GPT-3: Its nature, scope, limits, and consequences". In: *Minds and Machines*, pp. 1–14.

Fodor, Jerry A (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. The MIT Press.

Glock, Hans-Johann (2019). "Agency, Intelligence and Reasons in Animals". In: *Philosophy*, pp. 1–27.

Goldberg, Yoav (2016). "A primer on neural network models for natural language processing". In: *Journal of Artificial Intelligence Research* 57, pp. 345–420.

– (2017). "Neural network methods for natural language processing". In: *Synthesis Lectures on Human Language Technologies* 10.1, pp. 1–309.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Gubelmann, Reto and Siegfried Handschuh (2022). "Context Matters: A Pragmatic Study of PLMs' Negation Understanding". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4602–4621.

Harnad, Stevan (1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1-3, pp. 335–346.

Hassan, Hany et al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: *CoRR* abs/1803.05567. arXiv: 1803.05567. URL: http://arxiv.org/abs/1803.05567.

Haugeland, John (2002). "Syntax, Semantics, Physics". In: *Views into the Chinese Room*. Ed. by John Preston and John Mark Bishop. Oxford University Press, pp. 379–393.

Jacob, Pierre (2019). "Intentionality". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University.

Joshi, Mandar et al. (Nov. 2019). "BERT for Coreference Resolution: Baselines and Analysis". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5803–5808. DOI: 10.18653/v1/D19-1588. URL: https://aclanthology.org/D19-1588.

Kant, Immanuel (1999/1785). *Grundlegung zur Metaphysik der Sitten*. Felix Meiner.

Kassner, Nora and Hinrich Schütze (July 2020). "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7811–7818. DOI: 10.18653/v1/2020.acl-main.698. URL: https://aclanthology.org/2020.acl-main.698.

Khalid, Samia (Sept. 17, 2019). *BERT Explained: A Complete Guide with Theory and Tutorial*. URL: https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/ (visited on 05/22/2020).

Koehn, Philipp (2017). "Neural machine translation". In: *arXiv preprint arXiv:1709.07809*.

Läubli, Samuel, Rico Sennrich, and Martin Volk (2018). "Has machine translation achieved human parity? a case for document-level evaluation". In: *arXiv preprint arXiv:1808.07048*.

Li, Linyang et al. (2020). "Bert-attack: Adversarial attack against bert using bert". In: *arXiv preprint arXiv:2004.09984*.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Millikan, Ruth Garrett (1989). "Biosemantics". In: *The Journal of Philosophy* 86.6, pp. 281–297.

Moshman, David and Mark Timmons (1982). "The construction of logical necessity". In: *Human Development* 25.5, pp. 309–323.

Nagel, Thomas (1974). "What is it like to be a bat?" In: *The philosophical review* 83.4, pp. 435–450.

Oppy, Graham and David Dowe (2021). "The Turing Test". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University.

Peregrin, Jaroslav (2021). "Do Computers" Have Syntax, But No Semantics"?" In: *Minds and Machines*, pp. 1–17.

Preston, John and John Mark Bishop, eds. (2002). *Views into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press.

Proudfoot, Diane (2002). "Wittgenstein and the Chinese Room". In: *Views into the Chinese Room*. Ed. by John Preston and John Mark Bishop. Oxford University Press, pp. 167–180.

Quine, Willard Van Orman (1960). *Word and Object*. The MIT Press.

Radford, Alec, Karthik Narasimhan, et al. (2018). *Improving language understanding by generative pre-training*.

Radford, Alec, Jeffrey Wu, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (2018). "Know what you don't know: Unanswerable questions for SQuAD". In: *arXiv preprint arXiv:1806.03822*.

Rheinwald, Rosemarie (1992). "Das "Chinesische Zimmer" als Test des Turing-Tests? Zur Frage, ob Maschinen denken können". In: *Philosophische Rundschau*, pp. 133–156.

Roberts, Adam, Colin Raffel, and Noam Shazeer (2020). "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" In: *arXiv preprint arXiv:2002.08910*.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://aclanthology.org/2020.tacl-1.54.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.

Searle, John (1980). "Minds, brains, and programs". In: *Behavioral and brain sciences* 3.3, pp. 417–424.

– (1989). *Artificial Intelligence and the Chinese Room: An Exchange*.

– (1990). "Is the brain's mind a computer program?" In: *Scientific American* 262.1, pp. 25–31.

– (1993). "The problem of consciousness". In: *Consciousness and cognition* 2.4, pp. 310–319.

– (2007). "Putting Consciousness Back in the Brain". In: *Neuroscience and philosophy: Brain, mind, and language*. Columbia University Press, pp. 97–206.

– (2014). "What your computer can't know". In: *The New York review of books* 9.

Searle, John, S Willis, et al. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

– (1984). *Minds, brains, and science*. Harvard University Press.

Shanker, Stuart G. (1998). *Wittgenstein's Remarks on the Foundations of AI*. Taylor & Francis Group.

Sun, Ron (2014). "Connectionism and neural networks". In: *The Cambridge handbook of artificial intelligence*. Cambridge University Press, pp. 108–127.

Taylor, Charles (2016). *The language animal*. Harvard University Press.

Tsai, Yao-Hung Hubert et al. (2019). "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. Vol. 2019. NIH Public Access, p. 6558.

Turing, A. M. (Oct. 1950). "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. URL: https://doi.org/10.1093/mind/LIX.236.433.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Wang, Alex, Yada Pruksachatkun, et al. (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: *Advances in Neural Information Processing Systems*.

Wang, Alex, Amanpreet Singh, et al. (Nov. 2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://www.aclweb.org/anthology/W18-5446.

Widdows, Dominic (2004). *Geometry and Meaning*. Center for the Study of Language and Information, Stanford.

Wittgenstein, Ludwig (2006/1953). "Philosophische Untersuchungen". In: *Werkausgabe Band 1*. Suhrkamp.

Yang, Zhilin et al. (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems*, pp. 5753–5763.