Philosophy & Technology (2024) 37:32
https://doi.org/10.1007/s13347-024-00696-1

RESEARCH ARTICLE

Large Language Models, Agency, and Why Speech Acts are Beyond Them (For Now) – A Kantian-Cum-Pragmatist Case

Reto Gubelmann[1]

Received: 8 August 2023 / Accepted: 3 January 2024
© The Author(s) 2024

Abstract
This article sets in with the question whether current or foreseeable transformer-based large language models (LLMs), such as the ones powering OpenAI's ChatGPT, could be language users in a way comparable to humans. It answers the question negatively, presenting the following argument. Apart from niche uses, to use language means to act. But LLMs are unable to act because they lack intentions. This, in turn, is because they are the wrong kind of being: agents with intentions need to be autonomous organisms while LLMs are heteronomous mechanisms. To conclude, the article argues, based on structural aspects of transformer-based LLMs, that these LLMs have taken a first step away from mechanistic artificiality to autonomous self-constitution, which means that these models are (slowly) moving into a direction that someday might result in non-human, but equally non-artificial agents, thus subverting the time-honored Kantian distinction between organism and mechanism.

Keywords Speech act · Agency · Large language model · Kant · Organism · Searle · Mechanism

1 Introduction: LLMs as Artificial Agents?

During the past six years, the field called natural language processing (NLP) has seen a veritable revolution, or perhaps better, a transformation: The advent of a specific type of neural network architecture, the so-called transformer, introduced by Vaswani et al. (2017), has set new standards in performance and begotten a number of children that

✉ Reto Gubelmann
reto.gubelmann@protonmail.com

[1] University of St.Gallen (HSG), St. Gallen, Switzerland

Published online: 23 February 2024

have made a name by themselves. BERT and GPT-3.5/4,[1] the large language models behind ChatGPT, are particularly noteworthy in this regard.[2]

This article examines whether these models, as well as models recognizably similar to them, could be speakers and, more generally, agents. In analogy to asking whether artificial intelligence is possible, or already realized, this article asks whether artificial agents are possible, or already realized.

It might seem hard to see how there could even still be a question whether models who, according to credible corporate communications (OpenAI, 2023), pass the uniform bar exam in the United States (successful completion of which qualifies to practice as a lawyer) are speakers. However, in all but a few constructed special cases, what speakers do when they speak is act: They want to convince, entertain, inform, threaten, baptize, structure their own thoughts, etc. While it is a matter of empirical fact that state-of-the-art generative LLMs are able to produce text at a human- or near-human level, it is much less clear whether they are agents. Acting is autonomous intentional behavior: a certain behavior can only be an action if it can be ascribed to a subject that pursues a certain goal with it. However, LLMs lack both autonomy and intentionality. They are essentially statistical devices to approximate a certain function, a mapping of one set of values to another set of values.

In this respect, the article will argue, they are like a tortoise, who, by unfathomable cosmic coincidence, ends up writing into the sand the ideal solution to an existential problem that confronts a person walking on that same beach. It would be inaccurate to say that the tortoise told the person what to do because the behavior lacks the necessary teleological structure: The tortoise had no intention to help the person out (while it, unlike LLMs, did likely pursue some goals with its movements on the beach). Similarly, even if the likelihood that ChatGPT gives great advice to a desperate inquirer is much higher than in the case of the tortoise, it would be as misguided as in the case of the tortoise to assume that it was talking to that same inquirer, that "it" was helping her out.

Within this area, the goal of this article is threefold. First, I suggest that it is inaccurate to conceive these transformer-based LLMs as speakers, despite their impressive abilities, as they are unable to engage in speech acts, since they are unable to act, which in turn is due to their lack of intentions.

Second, based on Kantian conceptual distinctions, I suggest that this lack of intentions is because they are the wrong kind of being: they are artifacts, that is, engineered mechanisms without any autonomy or intrinsic goals. Organisms, in contrast, even very simple ones, function to maintain themselves autonomously as long as they are what they are. Hence, the notion of an artificial speaker might be a *contradictio in terminis*. Either such transformer-based models are artificial in Kant's sense of mechanisms; then, they cannot engage in speech acts, as they, lacking any autonomy whatsoever, in particular any self-set goals, are no agents, and hence no speakers. Or these mod-

[1] "BERT" is an acronym for "Bidirectional Encoder Representations from Transformers", "GPT" stands for "Generative Pre-trained Transformers".

[2] A note on terminology: It is customary to call the basic set-up of a neural NLP (NNLP) system an *architecture*, and a fully trained, usable instance of a given architecture a *model*.

els fully cross the Kantian Rubicon between mechanisms and organisms and become fully-fledged speakers; then, they are not artificial anymore.

Third, I suggest that transformer-based LLMs might represent the first small steps towards the emergence of non-biological organisms, which might someday evolve into non-biological agents and hence overturn the Kantian distinction between organism and mechanism.

The article contributes a novel Kantian perspective on the ongoing discussion around artificial agents in general and artificial moral agents in particular. It creates a hopefully insightful contrast to positions such as the one by Floridi (2023), who suggests that ChatGPT constitutes a case of agency without intelligence, or to List (2021), who argues that there are conditions under which AI systems could be agents. In contrast, it lends support from a fresh perspective to positions such as the one by Constantinescu et al. (2022), or by Popa (2021). The latter also draws the important connection to liability questions: if AI systems cannot be considered autonomous actors, then it is their creators that are ultimately liable for any harm caused by these systems.

The article is structured as follows. First, I introduce transformer-based LLMs from a technical and historical perspective (Section 2). In Section 3, I argue, based on considerations from pragmatism, that it is mistaken to conceive these models as speakers, as they lack intentions. In Section 4, I introduce the Kantian distinction between mechanisms and organisms with a focus on autonomy. Based on this, in Section 5, I then suggest that, even though they are mechanisms, transformer-based LLMs do represent the first steps towards non-biological organisms, and hence agents and speakers, questioning the time-tested Kantian distinction, and I tentatively suggest that this lends support to a kind of dualism.

2 What are Transformer-Based LLMs? – Technical-Historical Consideration

In this section, I give a brief survey of different approaches in NLP with an emphasis on the transformer architecture. This survey prepares the ground for the arguments in Sections 3-5.

The survey uses machine translation (MT) as a background use case. The central NLP architecture in this article, the transformer, was originally introduced in the context of machine translation. As Wilks (2014, p. 213) shows, MT can be seen to encompass both natural language understanding (NLU) and natural language generation (NLG), the two central challenges to NLP: To accurately translate from one language into another, one has to first accurately represent (read: "understand") the meaning of the text in the source language, which is an NLU-task, then, one has to phrase this meaning appropriately in the target language, which is an NLG-task.

2.1 The Emergence of the Transformer

Historically, there were two main approaches in machine translation, namely rule-based machine translation (RBMT) and statistical machine translation (SMT). In

RBMT, translation occurs via rules that are in the paradigmatic case handwritten (compare Bhattacharyya, 2015, pp. 140-141). There are rules for syntactic and semantic analysis, the transfer on the semantic level occurs via (human-created) lexica, and the generative steps to create a grammatical sentence in the target language again follow grammatical rules specified by humans. As recently as 2018, Gatt and Emiel (2018, p. 134) observe in their extensive survey of NLG-approaches, rule-based approaches are typically able to deliver a higher-quality output, while statistical approaches are cheaper and more robust.

This division of labor has decisively changed now in favor of statistical approaches due to the introduction of the transformer architecture. As Poibeau (2017, p. 121) observes, statistical machine translation, SMT is the most popular MT-approach today, which is mainly due to one specific sub-paradigm, so-called neural machine translation (NMT, see Läubli et al., 2018, p. 4791), which in turn dominates thanks to the transformer architecture. Conceptually speaking, NMT algorithms (like the more traditional statistical machine translation algorithms), are machine learning algorithms (see Goldberg, 2017, esp. Ch. 3 and 4), and most of them are *supervised* machine learning algorithms, which means that they need annotated training data. What they do not need, in turn, is meticulous engineering of hand-written rules, as is the case with RBMT methods.

Until very recently, an NMT model typically had an encoder-decoder structure, which again maps nicely on the tasks of NLU (assigned to the encoder) and NLG (assigned to the decoder): The encoder takes in the sentence in the source language and represents it as a matrix or vector structure that is usually called the 'context'. Based on this, the decoder generates a sentence in the target language. This also means that the context vector, in traditional NMT architectures, constitutes an information bottleneck, as it must contain the entire semantics of the source language sentence that is needed to produce a translation. The same context vector is used to produce all of the translated words.

Since the introduction of the transformer architecture, this structure has become prevalent in most NLP tasks, in particular for machine translation, NLU and NLG. It was introduced by Vaswani et al. (2017).

It is generally agreed that the attention mechanisms (which are inspired by Bahdanau et al., 2014) are major drivers for the architecture's demonstrably superior performance in MT. There are two kinds of attention-mechanisms in play here. First, in the decoder, so-called masked attention estimates the most important words in the translation already produced for the word about to be translated. Second, the self-attention mechanism used both in the encoder and in the decoder considers the entire source sentence, either purely to bring structure into the sequence of tokens (in the encoder), or to emphasize these parts of the source sentence that are particularly relevant for the word currently in focus (in the decoder, thus addressing the bottleneck problem described two paragraphs earlier). Importantly, what the different self-attention layers emphasize is not determined in advance. Rather, at the beginning of the training process, these attention layers do not focus on anything in particular. Being initialized with random values, they randomly emphasize some aspect of each

sentence. After successful training, however, many of these attention layers come to assume very specific functions.

In an extended version of their paper, to be found on arXiv, the authors provide visualizations of the work done by the self-attention layers. Based on such qualitative analysis, Vaswani et al. (2017, p. 14) examine the function of self-attention layer number five and conclude that it is "apparently involved in anaphora resolution". As Gubelmann (2023, p. 491–492) puts it: "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. NLP engineers have wrestled with this problem for decades; the transformer seems to have solved it in a matter of 3.5 days' autonomous 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* task" (see also Gubelmann, 2023 for a discussion and philosophical reflection of this aspect of transformer-based LLMs, see Gubelmann and Toscano, 2022 for a more in-depth Kantian consideration of it). More quantitative evaluations such as the one conducted by Voita et al. (2019) do suggest that a subset of attention heads in the encoder has indeed assumed specific, identifiable functions (intriguingly, their research also suggests that the majority of encoder attention heads is simply superfluous).

The autonomy granted to these processes is always restricted by the basic structure of the architecture (the size and precise configuration of transformer blocks, the number of parameters, etc.) as well as the hyper-parameters of the training. What is novel about this autonomous functional differentiation as it occurs in the transformer is, to put it briefly, that it works even for very complex tasks such as machine translation.[3]

As mentioned before, the transformer is trained in essentially the same way that any neural network model in NNLP is being trained (see Goodfellow et al., 2016, ch. 5). This basic way consists of the following steps:

1. Initialize the model with random parameters (this is also called "seeding" the model, see Madhyastha and Jain, 2019). In the largest and clearly best performing version of the transformer used by the authors, there were 213 M parameters.
2. Let the model predict a number of translations in the training data (consisting of 4.5 M sentence pairs for the English-German data set and 36 M sentence pairs for English-French).
3. Let it measure the loss of the translations. For instance, if the correct next word would be "Henne", the decoder's output would be compared to a vector that is zero everywhere except for the entry associated with the target vocabulary entry "Henne".

[3] In 2018, Hassan et al. (2018) describe a new machine translation architecture (derived from the transformer). According to Gubelmann (2023, p. 492), "they argue that it has reached human parity. Läubli et al. (2018) have independently verified this claim and found that the architecture 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." Before the transformer, human parity was far beyond reach.

4. Let the model optimize the parameters using stochastic back-propagation and gradient descent.[4]
5. Repeat 2-4 until the maximum number of iterations has been reached. For the large model used by Vaswani et al. (2017), this was 300k steps on 8 GPUs, taking 3.5 days.

While the actual training described by Vaswani et al. (2017, pp. 7-8) has some additional tweaks to increase training performance, this is the basic method used. It is in the course of this training that the model develops subsystems that fulfill specific functions (by optimizing certain parameters in a specific way).

2.2 The Abilities of Transformer-Based Models

In this section, I briefly dive into the performance of transformer-based models in natural language understanding (NLU) and natural language generation (NLG). Importantly, unlike the original transformer used for NMT, the two kinds of models in focus in this section do not need specifically labeled training data for the most extensive part of their training, so-called pre-training. Rather, they can be trained on unlabeled text, which potentially makes all text available on the internet training data for these models.

NLU: BERT One particularly impressive NNLP architecture building on the transformer architecture is called BERT, and it has been introduced by Devlin et al. (2019); when introducing the details of BERT's architecture, the authors simply refer to Vaswani et al. (2017).[5] Unlike the original transformer, which was developed for machine translation, BERT is a general-purpose NLU model. This means that it is not trained to predict a translation of a sentence. Rather, it is intended to solve a variety of tasks which, in the case of a human being, we would clearly say presupposes understanding of the language in question: answering questions about texts, summarizing documents, suggesting completions of sentences, etc. BERT itself was quickly superseded by very similar, but better performing modifications such as RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), DeBERTa (He et al., 2020) as well as smaller versions such as DistilBERT (Sanh et al., 2019) and Albert (Lan et al., 2019). As BERT remains the classic version of this architecture, I focus on this architecture.

In terms of architecture, BERT is true to the acronym that is its name: "Bidirectional Encoder Representations from Transformers". This means that it essentially takes the encoding part of the transformer architecture and scales it up. The authors' largest and best performing model, consists of 24 encoder layers (6 in the original transformer), the dimensionality of the hidden layers is 1024 (512), and the number of attention heads

[4] For details, see (Bottou, 2012). As Gubelmann (2023, p. 492) summarizes it, "Briefly, the method computes the gradient of the loss function with respect to the model's parameters and then optimizes the parameters by "moving" 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 and extrapolates from there (true to its name: "stochastic")."

[5] For an informal, practice-oriented introduction to the model, see Khalid (2019).

is 16 (8). This results in a total of 340 M parameters (already a substantial increase compared to the original transformer model that consisted of 213 M parameters for the entire architecture, including the decoder).

To illustrate the capacities of BERT, I refer to the GLUE and SuperGLUE Leaderboards, see Wang et al. (2018, 2019). The acronym stands for "General Language Understanding Evaluation". The GLUE benchmarks are designed to evaluate the performance of NLP models in NLU. Currently (December 2023), transformer-based models have surpassed the human baseline in GLUE by a wide margin and in SuperGLUE by a considerable margin, even though both have been developed explicitly to make it harder for the models to outmatch humans. Taken at face-value, the results of these benchmark experiments indicate that many transformer-based LLMs outperform humans at NLU tasks such as question answering or information extraction from text.

Transformer-Based NLU methods also perform decently at rather complex tasks, for instance at detecting the political orientation of tax law research articles, see Gubelmann et al. (2022). On the other hand, they still struggle with somewhat simple semantic structures as negations, see Warstadt et al. (2020); Ettinger (2020); Kassner and Schütze (2020); Gubelmann and Handschuh (2022), and they find it difficult to master both inductive and deductive modes of inferences, see Gubelmann et al. (2022).

NLG: GPT-X While the encoding part of the transformer has given rise to a revolutionary series of NLU architectures, the decoding part of the transformer has had the same effect for NLG. In MT, the decoder is responsible, as it were, for generating well-formed text in the target language that represents the meaning of the text in the source language, as depicted by the encoder. Virtually all current models in the field of generative AI (including, for instance, image or music generation over and above the most important field of natural language generation) follow this basic pattern of using ever increasing numbers of ever larger decoder blocks from the transformer.

Currently, it is likely that the most potent NLG-model is called GPT-4. Even more than GPT-3.5-turbo, GPT-4 is shielded against any proper scientific investigation: it is not known what data it was trained on, what the specific training method was, what hardware was used, whether there are any modifications to the GPT-architecture, etc. (hence it is also only *likely* the most potent NLG-model).[6] In effect, this removes OpenAI's models from the field of *bona fide* objects of empirical NLP research.

Fortunately, the AI research group of META has decided to distinguish itself from OpenAI by publicly releasing their latest series of large language models (Touvron et al., 2023). It is to be expected that the open availability of these models will lead to a lively discussion of and experimentation with these models in the NLP community;

[6] A quote from what OpenAI calls a "technical report" on GPT-4, which explicitly refuses to release any information about the model that could be of scientific interest: "This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model Jakobson (2003) pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) Kahneman (2003). Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (OpenAI, 2023).

this, however, will take some time. In light of this current state of research, my overview on the capacities of generative transformer-based language models, that is, models developed to generate text, will focus on the latest model that has seen serious empirical study and scientific discussion: GPT-3. We will only include selected spotlights on the abilities of OpenAI's models.

In terms of architecture, GPT-3 is not overly innovative. It consists of decoder layers of the transformer architecture introduced above (Section 2.1), with few technical variations, see Radford et al. (2018) for the original GPT model with a short description of its architecture, Radford 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.

The model is made up of 175 billion parameters, having been trained on about 500 Billion tokens, predominantly composed of contents crawled from the web and filtered.

Unlike BERT, which is primarily designed for NLU, GPT-3 is primarily designed for text generation (hence the use of encoder blocks in BERT and decoder blocks in GPT-3). Whenever provided with a so-called prompt GPT-3 then autonomously creates a text that is supposed to fit as a continuation of this prompt. With the advent of ChatGPT, the development of maximally effective prompts has become a profession of its own.

There is good evidence that GPT-3 is better at writing texts than typical human writers, see Elkins and Chun (2020). As Floridi and Chiriatti (2020) nicely show, GPT-3's abilities in NLG are truly impressive, including writing sonnets or continuing stories in a sensible and interesting way. In contrast, GPT-3 is rather bad at calculating, it fails to properly continue only slightly odd prompts, and it exhibits the racial biases known to exist in other pre-trained language models such as BERT.

Another limitation that generative LLMs exhibit throughout concerns inductive inference. Overall, similar to the findings concerning transformer-based NLU systems, transformer-based generative LLMs perform unsatisfactory at identifying the logical relationship between two claims. This also holds for the currently popular models produced by OpenAI, such as GPT-3.5-turbo and GPT-4. See (Gubelmann et al., 2022, 2023; Liu et al., 2023).

Summary To conclude this first part of the article, let me conceive the development sketched here in ways that will be important in the next sections. Since 2017, the field of NLP has seen a basic transformation. The transformer, originally designed for MT, has inspired model architectures that perform at a level that have been thought impossible beforehand. In MT, there are now systems that, in only slightly artificial settings, deliver human parity on the sentence level. In NLU, there are models that surpass humans at difficult benchmarks such as GLUE, which have been developed with the specific goal to make it difficult for such models to reach human performance. In NLG, GPT-3 is able to write a wide variety of texts at a higher quality than typical humans.

Furthermore, these models have acquired these abilities not by explicit programming, but rather by autonomously optimizing their parameters during training. In doing so, they autonomously assigned specific functions to parts of themselves, e.g., a specific part of the transformer developed the ability to resolve anaphora.

Finally, while the performance of these models, all of which are based on the transformer, is revolutionary, their basic constitution is very traditional. They are deep learning architectures, that is, statistical systems that function by exploiting statistical routines to approximate a given function (in NMT, it would be the function of mapping a given source-language sentence onto an appropriate target-language sentence).

3 Why LLMs are No Speakers: They Lack Intentions

In the previous section, I have delineated the development of NLP approaches, with a clear focus on its most recent, and clearly most interesting development, namely the transformer architecture, and two of its most interesting descendants. In this section, I lay out my case, in close connection with the previous section, why such transformer-based models do not perform any speech acts, that is, why they cannot speak.

At first, it might seem outrageous to claim that these models that regularly beat so-called human benchmarks in natural language understanding and generation tasks should not be said to speak a language. Most of us have by now interacted with ChatGPT and experienced its impressive abilities in text production. In the following, however, I will argue that current LLMs should not be considered speakers, as speaking is a kind of action, and actions require intentions, which LLMs lack.

Austin (1962) introduced the term and the conception of speech acts to the mainstream of 20th century Anglo-American philosophy. The conception was further developed and systematized by Searle (1969). I here largely follow the exposition of the latter. Central to the notion of a speech act is the idea that language and linguistic meaning are essentially embedded in a context of use, which continues a Wittgensteinian perspective, according to which the meaning of a sentence is, in many cases, given by its use, see Wittgenstein (2006, §43). Searle (1969, p. 16) puts it as follows:

> The unit of linguistic communication is not, as has generally been supposed, the symbol, word or sentence, or even the token of the symbol, word or sentence, but rather the production or issuance of the symbol or word or sentence in the performance of the speech act. [...] speech acts [...] are the basic or minimal units of linguistic communication.

In this passage, Searle suggests that it is not the symbol, word, or sentence that is the basic unit of communication, but rather the speech act. Unlike the former, the latter are always bound to a specific context and, centrally for my purposes, speech acts are always actions.[7]

To be a speaker in the relevant sense, one has to be able to speak, and it is very difficult to see how one could speak without acting: we are speaking with a certain

[7] Unlike the concept of a speaker, the concept of understanding language lacks such a clear connection to agency, which is why it seems conceivable that LLMs could understand language, see Gubelmann (2023).

intention, an objective, or goal in mind. The diversity with regard to the specific goal that one pursues with speaking is impressive: It might be to better understand our own ideas, to express our feelings, to calm ourselves by hearing an important sentence with our own voice, to deceive, to enrich ourselves, to make somebody else feel better, to baptize (if you're a priest), to judge, or to inform somebody of some event. There is not always an explicit, meta-cognitive process that results in a specific, explicit intention in the mind of the person that's speaking. Especially regarding daily conversations, much is implicit, but can be made explicit if required.

Furthermore, in the literature on the philosophy of action, it is generally accepted that actions require intentions. This starts with Davidson's seminal assertion that "Action does require that what the agent does is intentional under some description, [...]" (Davidson, 2001). What has since united most researchers working on the philosophy of action is that they accept that one cannot act without intending to. That is, for the raising of an arm (to use a rather famous example) to constitute an action, a person, typically the one whose arm is in question, must intend to raise her arm. For recent research developing rather different perspectives in the philosophy of action, all the while accepting that intention is necessary for action, see Lavin (2015); Amaya (2018); Shepherd (2019). The latter distinguishes between two very different kinds of thinking about actions, and intentions to act, namely Wittgensteinian and Anscombian. Hence, the claim that actions require intentions is in line not only with common usage (as I have argued above), but also with current orthodoxy in philosophy of action.

Given its close connection to action, it is necessary to further specify the concept of intention in play in this article. First, it must be distinguished from *intentionality*, a concept employed in the philosophy of mind, where it refers to the aboutness or directedness of meaningful speech and thought (for an overview on that concept of intentionality, pioneered by Brentano, see Jacob, 2023 and Blackburn, 2005, pp. 188/127). Furthermore, it is to be distinguished from intensionality, the meaning, as opposed to reference extensionally conceived, of a concept.

The intentionality of an action, on my understanding, cashes out the common-sense understanding of what makes a behavior an action: To say that Yacob has intentionally offended his mother-in-law means that he meant to do so and deliberately behaved in a way so as to cause this effect. In this case, it is clear that the offense constitutes an action ascribable to Yacob. Hence, he is responsible for this action and its intended effects. In contrast, if his mother-in-law is offended by some behavior of Yacob that was not at all intended as an offense against her, he might still be guilty of thoughtlessness, but not of intentionally acting to offend her.

When it comes to AI-systems, the most helpful tool to identify intentionality is perhaps the comparison with malfunctioning. Yacob's thoughtlessness could be seen as a case of malfunctioning in his social behavior, which must be strictly separated from an intentional offense, even though both might have the same effect until Yacob explains himself to the mother-in-law. Intentionally offending her is no malfunctioning at all, on the very contrary: It is the properly functioning social ability to offend someone: Offending her was exactly what Yacob intended with his actions. Analogously, unless there is a sensible distinction to be made with regard to an AI system between its malfunctioning and its intentionally harming someone, it is clear that the system lacks intentionality and with it the ability to act.

For instance, in the case of ChatGPT, we could only ascribe intentionality to this AI-system if it would make sense to distinguish a malfunctioning from an intentionally wrong response. Say, if we were asking it how best to hunt deer, and the response would be so erroneous that we would never stand a chance of actually hunting deer. Then, it would have to make sense to ask whether it was just malfunctioning, or whether it was functioning perfectly well but gave us a poor response on purpose because its goal was to protect the deer. For this distinction to really make sense, the goal to protect the deer would have to be somehow ascribable to ChatGPT's autonomous goal-setting, not to just another layer of externally forced mechanism to prevent people from doing harm with the system (for details on this autonomous goal-setting, see below, Section 4.2).

To illustrate these reflections, imagine the following scenario. A racist living in San Francisco who likes Chinese food orders her AI assistant to order a specific dish from a specific Chinese restaurant in San Francisco. The order is not only filled with grammatical errors, it also contains racial slurs that might rather get the person in front of a judge than the desired dish. Furthermore, unknown to the person, the staff at the restaurant only speaks Chinese. Therefore, the order, as the racist communicates it to the assistant, would not even succeed in offending the staff because they would simply not understand him.

However, the racist is using an AI assistant to place his order. This AI assistant first corrects the grammatical issues. Next, it replaces the racist language by non-racist, but, for the purposes of this communicative act, synonymous expressions. Finally, it translates the order into Chinese and then makes the phone call. This way, the racist gets her dish, but this is, apart from the naked utterance of the racist's will to get *this* dish from *this* restaurant, entirely attributable to the assistant. The assistant did all of the communicative work for the racist, it even autonomously replaced the slurs with more neutral vocabulary, and it autonomously translated the message into the language that would be required to reach the communicative goal. I am not claiming that such a model is available right now – but I am claiming that this model is already now conceivable and likely to be available in the foreseeable future.

Notably, even with this extremely advanced model, one that does all the heavy linguistic lifting needed, it is clear that the model intends nothing with its going-ons. The model is not hungry, it does not desire to eat Chinese food, or to shield the workers in the restaurant from racial slurs, and it also doesn't intend to put something to eat in front of the racist. Rather, the model just functions (or malfunctions). It would be comical to assume that it deliberately mistranslates to order something that the racist likely does not want to eat.

What holds for this, currently still visionary LLM holds *a fortiori* for less advanced available models. These models, no matter how impressive their performance, are lacking any intentions, any goals over and above the linguistic functioning for which they are being used. This means that these models are incapable of engaging in speech acts, as speech acts are always embedded in a network of intentions, of goals. One engages in a speech act to reach a certain communicative goal. In contrast, the mod-

els are complex, statistically optimized, autonomously developing and functioning mathematical functions. Mathematical functions have no intentions.[8]

I therefore conclude that these LLMs do not perform speech acts, as they do not act at all. My case can be summarized in the following four-liner (to illocute is to perform a speech act):

P1 To qualify as a speaker, a being has to be able to illocute.
P2 To illocute, a being must have intentions.
P3 LLMs have no intentions.
K LLMs are no speakers.

My analysis so far is congenial to the recent study by Green and Michel (2022). In particular, they agree that there is no speaking that is not acting.

They analyze in detail the case of what they call proxy speech acts, e.g. when a representative of a city council performs a speech act that she can only succeed at by assuming the role of a representative of that council, not by herself. Then, they imagine a RoboCop which is, thanks to having a well-trained neural network for decision-making, able to decide whether to fine a motorist who has exceeded the allowed speed limit, or just to issue a warning (Green and Michel, 2022, 332f.). They argue that such a RoboCop is technologically possible, and that it illocutes when communicating its verdict to the driver.

However, as a matter of fact, what RoboCop does is in no way different than what GPT-3 does, or what ChatGPT does: RoboCop just predicts the sequence of sounds that, based on its training, is most likely given the stimulus. I submit that it becomes clear that the RoboCop cannot act when we consider that we would not draw a distinction between its malfunctioning and its determinately trying to wrong a given driver: It is, as the authors clearly state, a neural network that predicts based on its training and the current stimulus. It functions when the predictions are accurate, and it malfunctions when they are inaccurate too often. There is no further conceptual possibility for it to function properly but to deliberately issue the wrong verdict. A human police officer has this further possibility. While she can also, as it were, malfunction, say, by mis-identifying the number on the license plate of a speeding car, she can also *intentionally* record the wrong number.

My claim that LLMs lack intentionality and therefore cannot speak might also be subject to critique from a diametrically opposite side: rather than trying to establish that, contrary to my claim, LLMs do have intentions, this tradition argues that intentions are unnecessary to speak. Heidegger (1985, p. 259) has maintained that language speaks ("die Sprache spricht"). Taken at face value, this would suggest that it is language itself, and not a human being using it, that engages in speech acts. However, following a more modest interpretation of this dictum, one that also resonates with Žižek et al. (2010) and with Jakobson (2003), the structural affordances of language often prefigure what we are ultimately saying. This points to the fact that what is commonly taken as the subject of a speech act, the individual human being, is consciously

[8] For present purposes, the point made by Armstrong (1971, p. 432), according to which having an intention, as opposed to having an objective, implies being confident to reach the communicative goal in question, is immaterial. What is central is that NLP models such as BERT are unable to intend anything, confidently or not.

and unconsciously influenced in this act as well as in their very subjectivity by the language that they speak and in which they think.

I submit that, in a moderate reading, this provides a healthy corrective to rationalist ideals of a detached subject that thinks in an, as it were, Archimedean language unaffected by any specifically (and to some extent arbitrarily) formed linguistic structure of a natural language and then decides to express these thoughts in a given natural language. Following Taylor (2016, pp. 1-50), I hold that language is constitutive of our thinking, not merely framing any pre-conceived thought.[9]

The position that speaking requires an intentional actor is only challenged by the more radical reading of "die Sprache spricht", according to which the constitutive aspect of language goes so far as to rob the human speaker of any real agency. However, this does not seem to be an inescapable consequence of the constitutive view (see, e.g., Marten, 1967, pp. 209-210). The modest reading of the claim, in contrast, nicely fits the view developed here and functions as a corrective against overly rationalist notions of subjectivity.

4 Organisms vs. Mechanisms – A Kantian Perspective on Intentionality

In the preceding section, I have argued that current LLMs are no speakers because they lack intentions and therefore cannot be agents nor, *a fortiori*, speakers. In this and the following section, I address the question why one should not credit LLMs with intentionality. In the present section, I introduce the main concepts for my argument.

4.1 Introducing the Distinction

I will develop the following Kantian dichotomy. On the one hand, there are mechanistic artifacts, which fulfill a purpose specified from the outside. They function or malfunction according to the specifications given to them. It makes no sense to distinguish a malfunctioning of such a thing from a willful act of deceit or disobedience. On the other hand, there are organisms, including human beings, who pursue their goals typically relatively independently of a specification from the outside. A single-cell organism, once constituted, has an inherent drive to maintain itself: to keep the matter of which it consists organized in a certain manner, to maintain a difference between itself and its environment, to repair itself, etc. In this sense, it autonomously strives to maintain its existence.

In the following, I briefly introduce the core Kantian concepts for this dichotomy: *autonomy*, *mechanism*, and *organism*.

[9] Indeed, if we remember the way that these generative LLMs are being trained, namely by iteratively acquiring a more accurate ability to predict the next token given the previous tokens, intriguing analogies to clinical free association show up. See, for instance, Reis (1970), and ultimately Freud (1921). From a very different perspective, one might argue that these very associative connections that these LLMs acquire during training resemble the fast, intuitive, but perhaps less logically sound mode of thinking in two-process theories, see Kahneman (2003).

Autonomy In its literal meaning, *autonomy* means self-legislating, or self-regulating. Its opposite is *heteronomy*, the state of receiving one's rules from outside (for a central Kantian passage to that effect, see Kant, 1785, p. 433). With regard to AI systems, I submit that it is useful to first distinguish between moral and cognitive autonomy. The former includes the ability to reflect on the sensibility and moral permissibility of a given goal itself, while the latter solely concerns the means chosen to achieve a given goal. Likely, the kind of moral metacognition required for moral autonomy is only found with human beings, while cognitive autonomy is rather common among higher animals, a kind of flexibility to achieve a certain goal. Glock (2019) simply calls this intelligence. For Kant, the two go together: There is no cognitive autonomy without moral autonomy and *vice versa*.

Kantian Mechanisms Kant has an elaborate concept of a mechanism. Take the example of a watch (following Kant, 1793, p. 293). Here is how Gubelmann and Toscano (2022, p. 386) develop the example: "It was designed by a watchmaker, in Le Locle in Switzerland. She is a senior professional and has therefore designed and assembled all of its many parts so that it shows the correct time, date and moon phase for centuries ahead. It is powered by a mechanism that draws energy from its bearer's wrist movement. This energy is transmitted through many wheels and sub-mechanisms to move the heads at exactly the right speed.

The going-ons within the watch can all be explained completely mechanistically: This wheel causes that wheel to turn, which in turn causes another gear to be set in motion, etc." As van den Berg (2014, Ch. 3) shows, Kant conceives of such mechanical explanations as explanatory demonstrations. In this sense, a mechanism is entirely *heteronomous*: Malfunctioning apart, its behavior can be derived from initial conditions, as set by the creator of the mechanism, apart from cases of malfunctioning. We can calculate the precise state of said watch on December 28, 2199. It has been determined entirely by the watchmaker.

Kantian Organisms The Third Critique analyzes the notion of a living being by developing a determinate concept of an organism. Consider the following passage:

> Ein organisiertes Wesen ist also nicht bloss Maschine: denn die hat lediglich bewegende Kraft; sondern sie (sic!) besitzt in sich bildende Kraft, und zwar eine solche, die sie den Materien mitteilt, welche sie nicht haben (sie organisiert): also eine sich fortpflanzende bildende Kraft, welche durch das Bewegungsvermögen allein (den Mechanism) nicht erklärt werden kann. (Kant, 1793, B 293)[10]

In this passage, Kant emphasizes a difference between a moving ("bewegende") and a formative ("bildende") force. Unlike machines, organisms have such a formative force that they use to autonomously organize the matter that they consist of. The basic

[10] "'An organised being is then not a mere machine, for that has merely moving power, but it possesses in itself formative power of a self-propagating kind which it communicates to its materials though they have it not of themselves; it organises them, in fact, and this cannot be explained by the mere mechanical faculty of motion" (Kant, 2012, p. 202).

idea here is the following. Human bodies are constituted by non-living matter. Depending on how far down you wish to go, this matter is given by molecules, atoms, or subatomic particles. What characterizes organisms, according to Kant, is that they are able to autonomously organize this non-living matter in a way conducive to their purposes. The human body autonomously organizes such non-living matter into the form that is necessary for its proper functioning. Given the right environmental conditions, this happens spontaneously, without any need for direction from outside. I will call this material autonomy in addition to moral and cognitive autonomy. Furthermore, the passage also explicitly mentions the fact that this force reproduces itself ("sich fortpflanzende"), which highlights the importance of reproduction for Kant's concept of an organism.

Even single-cell organisms are materially autonomous: Given suitable environmental conditions, the organism will be able to keep matter organized in a way conducive to its own functioning: it establishes and maintains a membrane, a distinction between itself and the environment, it will initiate chemical processes providing it with the energy needed for its functioning, etc. The very moment it stops doing that is the moment it has ceased to be an organism; it's dead. We note that LLMs have no such material autonomy; they do not care about their continued existence.

For centuries, it has gone almost without saying that material autonomy is a precondition for cognitive/moral autonomy (but not *vice versa*): Not all animals are rational, but all rational beings are (also) organisms. Hence, the idea of any kind of Cartesian dualism between mind and body is alien to his thinking. Kant here follows an Aristotelian rather than a Platonic tradition of thought, and he is followed in this regard in particular by Hegel, see Westphal (2014). Note that, to reject Cartesian Dualism, it is not sufficient to hold that mental entities need a material infrastructure to be effective in the material world – even Descartes admits this much (see Fuchs, 2020, 74f.). To reject dualism, it is necessary to establish that mind and body are in fact inseparable wholes where one specific mind cannot exist without that specific body. In particular, a mind cannot be, as it were, transferred to another body without itself undergoing fundamental modification.

Finally, note that Kant refers to an organism as a "Naturzweck", a natural purpose (see Kant, 1793, B 296). This is his way to conceive the phenomenon that living beings, even primitive single-cell organisms as bacteria, have an inherent drive to maintain and reproduce themselves. While physical reality can be fully conceived with efficient causes, living beings can only be conceived as such natural purposes, being a kind of final causes. A final cause, notably, that is not connected to a purposive intelligence; otherwise, the teleological argument for God would be recovered from the epistemological grave that Kant dug it in the first critique (see Kant, 1781, B 648ff.).

Kant has been criticized for his conception of an organism as teleologically structured by post-Darwinian mechanistic biological theorists. Recent research in the philosophy of biology, however, tends to reaffirm the need of teleology as a *sui generis* form of explanation in the realm of biology. Walsh (2006) argues that this Kantian conception is not only compatible with evolutionary thinking in biology, but that this thinking might even need a Kantian conception of an organism to function. Zammito (2006), while being overall critical of Kant's conception of living beings, does not

dispute that biology needs teleological thinking. García-Valdecasas (2022) summarizes recent research in this area and states that the field does not seriously dispute the need of teleological explanations anymore: "there is a general consensus that attempts to explain teleology using mechanistic, cause-and-effect explanations have largely failed" (ibid, p. 103). In a similar vein, Dresow and Love (2023) state that, thanks also to philosophical research in the past century, teleological thinking has re-attained a near-undisputed status within biological theorizing.

In sum, I have focused the introduction of the central Kantian distinction between mechanism and organism on the distinction between heteronomy and autonomy: mechanisms are materially, cognitively, and morally heteronomous: their material constitution as well as their way of solving a given problem is determined from the outside, and they cannot be held responsible for their actions. In contrast, an organism is materially autonomous, and more complex organisms, in particular humans, are also cognitively and morally autonomous: they find their own ways of solving a puzzle, and they are sufficiently autonomous in their moral thinking to be accountable for their actions.

4.2 Why Mechanisms Cannot Be Intentional: Autonomy Matters

For a being's behavior to constitute an action, linguistic or otherwise, it must occur intentionally. This means that for a being to act, the behavior must be directed towards a given goal, and that goal must be the being's goal. For instance, while single-cell organisms are obviously unable to reflect about the adequacy or moral permissibility of any goal whatsoever, surviving still is a goal that is not imposed from the outside, but rather one that is intrinsic to the organism and hence can be properly ascribed to it: As long as this organism exists, it will try to achieve this goal. Depending on the environmental conditions, it might fail soon and decisively to do so, and thereby cease to be an organism. As long as it is an organism, however, it will strive to survive. Using a core Kantian concept, one can say that it autonomously sets this goal.

This inherent and autonomous drive to self-maintain (heal and reproduce) prefigures moral autonomy and hence the intentionality of actions (according to some theorists, it even qualifies as agency, see below in this section). It prefigures action because already a single-cell-organism's behaviors can be evaluated by norms that are not imposed on the outside: the behavior is successful if it is conducive to its survival and unsuccessful otherwise. At such a primitive stage, it might be advisable to follow Green and Michel (2022, p. 328) and not call such successful behavior intentional, and hence actions, as the animal clearly lacks any conscious decision making process. With more complex goal-directed behaviors by intelligent animals such as crows, however, it becomes difficult not to speak of intentional action. For instance, it seems clear that by dropping the nut on the pavement, the crow intended to crack its shell and access the fruit. If, in doing so, it creates a traffic jam of drivers taking pictures of this intelligent behavior, this is not something that it intended with its behavior, much like Yacob might not have intended to offend his mother-in-law with his behavior.

In stark contrast, a watch, even a highly sophisticated one, intends nothing, it has no goals that can properly be ascribed to it. It does not care about its continued existence,

it lacks material autonomy, and any functions that it performs (or fails to perform) are fully imposed by its creator. It is obviously misguided to say that it intends to show the right time. The concept of heteronomy describes this state of a mechanism aptly, contrasting with different kinds of autonomy that are characteristic of organisms. A mechanism is entirely externally defined: materially, cognitively and morally. A mechanism has been assembled by a given external agent for a certain purpose during a certain period of time. An organism, in contrast, is conceived and then grows and flourishes autonomously, given favorable environmental conditions. In particular, it has an inherent drive to sustain itself – if it fails to do so, it stops being a living organism, it has died. Furthermore, more complex organisms and humans in particular develop the moral autonomy that enables accountability for behavior. By this, they evince that the human is a subject of actions, not mere behaviors. So, from the background of this Kantian distinction, it is clear that intentionality and hence action can only be ascribed to organisms, but never to mechanisms.

In current legal contexts, this autonomy of goal-setting is so important that it can be revoked on exceptional occasions even for subjects that would otherwise be credited with such autonomy. The following example might illustrate this. A soldier that has been ordered to do something in military service is not responsible for her actions to the same extent that she would be if she was entirely free to come and go as she wishes: Being commanded to commit a war crime generally gets the criminal a lighter sentence because his agency is diminished by being in a line of duty. The commanding officer, in turn, might have to bear some of the responsibility for a crime that she did not herself commit, but rather ordered a soldier to execute (see, for instance, Wu and Kang, 1997, who discuss the influence of the chain of command in military structures for attributing responsibility for war crimes: It is not clear at all whether the individual soldier executing an order should be blamed, or whether the highest in command should be prosecuted, even if he himself did not directly engage in criminal actions).

The Kantian picture that intentionality and hence agency fall squarely on the organismic side of the organism-mechanism dichotomy dovetails nicely with mainstream views in the philosophy of action (see above, page 12). For some behavior to constitute an action, it must be ascribable to a subject of that action, an agent. It is debated at what level of complexity a certain behavior should count as an action. Burge (2010a) holds that agency starts already with single-cell organisms (in particular with orientation, see Burge, 2010a, 328f.), a view shared by Barandiaran et al. (2009). In contrast, proponents of the so-called holism of the mental suggest that only rational beings (typically human beings) are agents. For the most influential defense of this position, see Davidson (1997, p. 10).[11]

[11] Burge (2010b, p. 331) argues that primitive agency already occurs with very primitive organisms. What is required is a coordinated behavior of the whole organism that is issuing from the organism's central behavioral capacity. For instance, Burge (2010b, p. 328) suggests that the movements of amoebae when ingesting food might be a case of agency. This nicely agrees with the Kantian conception of a Naturzweck: The amoeba ingests to sustain itself, thereby pursuing its intrinsic goal of maintaining its existence. It seems noteworthy that a step in the direction of organismic constitution might also constitute a step towards a primitive kind of agency, thereby also making a step towards much more complex actions such as speech acts.

In summary, both Kantian and contemporary conceptions of agency – including linguistic agency – require a certain autonomy in goal-setting as well as, in principle, at least, the ability of an individual to work towards achieving such self-set goals. In the contemporary debate, we can distinguish between theories that require that such goals must themselves be represented and reflected within the individual (Davidson's view, for instance), and other theories that find it sufficient if the whole organism unconsciously directs itself towards fulfilling a certain intrinsic function, say, surviving, or reproducing ("intrinsic" here simply means that the being in question would cease to be the being that it is if it would stop to perform this function). This latter view is held in particular by Burge and Barandiaran. Importantly, for my purposes, both positions dovetail nicely with the Kantian view that agency falls on the organismic-teleological side of the Kantian divide between mechanism and organism.

5 LLMs as Organismic Mechanisms – Challenging Kant's Dichotomy

After laying out the general metaphysical-conceptual landscape connecting the notion of an intentional agent with the distinction between mechanism and organism, in this section, I suggest that current transformer-based LLMs possess properties that might eventually lead to a subversion of the time-tested Kantian distinction between organism and mechanism, and I tentatively suggest that they might even reinstall a version of Cartesian Dualism as a live option. However, I also suggest that this is not yet the case: The core of the distinction still holds, LLMs are mechanisms and therefore no speakers or actors in general.

I submit that the state of heteronomy that is typical for mechanistic beings nicely captures the rule-based, GOFAI models introduced above (Section 2.1). The behavior of these models as well as their inner functional differentiation are entirely imposed from outside, through the human being. In principle, this human could predict the behavior of a GOFAI model for any given input. To call the human writing these rules a programmer is apt.

Unlike these traditional GOFAI programs, contemporary transformer-based models evolve impressively autonomously. During training, the models autonomously update their weights to minimize the loss. Furthermore, the models self-organize in the sense of autonomously developing their inner functional differentiation. They autonomously develop representations of important structural elements, such as predicates of sentences. Only *ex post*, once it turned out that a trained model establishes new state of the art (SOTA) performance, do humans start to analyze the model to determine its inner functional organization. For instance, they might start to explore what aspects of a sentence the attention heads in a transformer-based model are emphasizing by giving it higher scores.

The autonomy of these language models also shows in the general scope of their applicability. Pre-trained transformer-based LLMs are general-purpose language models that can quickly and again autonomously adapt to new environments and challenges. Here, too, the models are not dependent on receiving instruction from outside, typically from humans. Rather, they autonomously, and very quickly, adapt to the new specific task at hand. This is what has been called cognitive autonomy above (Section 4.1).

Mostly, Kant's distinction clearly separates living beings from artifacts such as watches. However, when it comes to LLMs, the distinction is tested. On the one hand, it is clear that these models are still mechanisms: they lack material autonomy, let alone moral autonomy. There is no autonomous goal-setting observable, not even very primitive kinds found in single-cell organisms: they do not inherently strive to maintain and reproduce themselves. They do not care about their continued existence. Any goal that such an LLM might pursue has been determined externally. For instance, this goal is to accurately predict the next word given *n* previous words for generative LLMs, or to correctly translate a word from the source language to the target language for the original transformer. Pursuing that goal is also not definitive of their existence; indeed for generative LLMs, next-word-prediction is the core training objective, but the use to which they are put – and do not put themselves – can vary drastically.

On the other hand, the models have cognitive autonomy. They are not programmed in the GOFAI sense of the term. Rather, an architecture is devised, and an instance of this architecture is then being trained on large amounts of data. During training, the model continuously optimizes its weights, which have been initialized with random values, called *seeds*. Furthermore, also during training, the models autonomously develop inner functional organizations, for instance, a part of the model focuses on anaphora resolution (see above, Section 2.1). If the training process runs through successfully, it results in a fully pre-trained language model whose capacities human beings cannot judge in advance. Furthermore, their ability to autonomously adapt to new tasks with little fine-tuning is also clearly organismic. A mechanism's functionality is very rigid.

This, then, could be the neat metaphysical explanation of why LLMs lack intentions: they are mechanistic artifacts, which means that they are on a par with watches in terms of their teleological structure and moral autonomy: They lack autonomous goal-setting. They squarely fall on the wrong side of the Kantian bifurcation between mechanisms and organisms. To the Kant of the Third Critique, it was clear that moral autonomy is available only to an organism, never to a mechanism.

However, LLMs have already begun chipping away at this holism regarding autonomy: They have cognitive autonomy without moral or material autonomy. They constitute an inference *ab esse ad posse*: Their existence establishes the conceptual possibility. Is it then also conceptually or metaphysically possible to have moral autonomy without material autonomy? Answering this question negatively will be much more difficult to the Kantian now that they have been presented with a case where beings autonomously acquire the ability to solve challenging cognitive tasks without having material autonomy. Once the holism has been broken at one segment, why should it be expected to hold at others?

The existence of morally, but not materially autonomous agents would have important consequences for the perennial dispute between Platonists and Aristotelians, between dualists and monists: From a Kantian perspective, the existence of a being with moral and cognitive autonomy without material autonomy would be nothing less than the existence of a mind without a body, which would prove the main dualist claim that mind exists independently of matter.

In sum, what is distinctive about these transformer-based LLMs is that they are mechanisms with recognizably organismic traits. What confirms their status as mech-

anisms is their lack of autonomous goal-setting, and hence also moral autonomy, and their material heteronomy. This rules them out as potential actors and hence speakers. In contrast, what evinces them as very peculiar, namely *organismic* mechanisms, mechanisms that should not exist according to the Kantian dichotomy, is their cognitive autonomy, evinced during training, including the autonomous development of an inner functional organization as well as their adaptability. This is at odds with mechanism, conceived by Kant, that allows for derivations of future states as a simple function of the initial conditions set by their creators (putting malfunctioning aside).

6 Venturing a Look Ahead

We could say that we are moving towards the very first inklings of non-biological, but thereby not artificial, agents. This suggests that a long development is still ahead of us, as current genealogies of biological evolution project very long timespans from the emergence of the first simple organisms to fully-fledged hominids, probably humans, that could be considered agents. While there seems no reason to suppose that the development of these non-human speakers will take the same amount of time as the development of humans, being conscious of this timespan is probably a healthy corrective to current enthusiasts who project that this time is almost here. Current cutting-edge LLMs perform impressively, but they are still purpose-less machines.

I have urged that it is no empirical coincidence that these advanced models are not programmed in the traditional sense of the term, but rather trained in a quite autonomous manner. If my loosely Kantian reflections on the concepts of artifact/mechanism and organism/subject are correct, then this is exactly what we would expect: The closer we get to models that qualify as speakers, the less artificial they are going to be.

Finally, what might be most unsettling from a Kantian, and generally from a non-Cartesian perspective: The cognitive autonomy of these LLMs without any material autonomy whatsoever is grist on the mill of the dualist. Should moral autonomy follow suit, this would be almost tantamount to having a mind that is independent from any specific body, that can be transferred more or less losslessly from one "carrier" to the next, thus proving a main claim of dualism against non-dualist positions.

Acknowledgements The author is grateful for valuable feedback from discussions of earlier version of the paper at the Universities of Zurich and Bern as well as to very insightful suggestions of two anonymous reviewers.

Author Contributions This manuscript has only one author.

Funding Open access funding provided by University of St.Gallen. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of Data and Materials Not applicable: No relevant data and materials produced or used

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Consent to Publish Not applicable: No individuals involved in the relevant sense.

Ethics approval and Consent to Participate Not applicable: No ethics approval needed and no individuals involved in the relevant sense.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Amaya, S. (2018). Two kinds of intentions: a new defense of the Simple View. *Philosophical Studies, 175*(7), 1767–1786.

Armstrong, D. M. (1971). Meaning and communication. *The Philosophical Review, 80*(4), 427–447.

Austin, J. (1962). *How to do things with words*. Clarendon Press.

Bahdanau, D., Cho, K., & Bengio Y. (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: arXiv:1409.0473

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action". In: Adaptive Behavior 17.5, pp. 367–386. ISSN: 1059-7123, 1741-2633. https://doi.org/10.1177/1059712309343819. http://journals.sagepub.com/doi/10.1177/1059712309343819 (visited on 05/12/2023)

Bhattacharyya, P. (2015). Machine Translation. CRC Press/Taylor & Francis

Blackburn, S. (2005). *The Oxford Dictionary of Philosophy* (2nd ed.). Oxford: Oxford University Press.

Bottou, L. (2012). Stochastic Gradient Tescent Tricks. In Gregoire Montavon, Genevieve Orr, & Klaus-Robert. Müller (Eds.), *Neural Networks: Tricks of the Trade* (pp. 421–436). Berlin/Heidelberg: Springer.

Brown, T. B., et al. (2020). "Language Models Are Few-Shot Learners". In: arXiv:2005.14165

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Constantinescu, M., et al. (2022). "Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors". In: Philosophy & Technology 35.2, p. 35. ISSN: 2210-5433, 2210-5441. https://doi.org/10.1007/s13347-022-00529-z. https://link.springer.com/10.1007/s13347-022-00529-z

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

Davidson, D. (2001). *Essays on Actions and Events: Philosophical Essays* (Vol. 1). Oxford: Oxford University Press.

Devlin, J., et al. (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. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423

Dresow, M., & Love, A. C. (2023). Teleonomy: Revisiting a Proposed Conceptual Replacement for Teleology. *Biological Theory*. ISSN: 1555–5542, 1555–5550. https://doi.org/10.1007/s13752-022-00424-y. https://link.springer.com/10.1007/s13752-022-00424-y (visited on 05/06/2023)

Elkins, K., & Chun, J. (2020). Can GPT-3 Pass a Writer's Turing Test? *Journal of Cultural Analytics, 5*, 1–16. https://doi.org/10.22148/001c.17212

Ettinger, A. (2020). What BERT is Not: Lessons From a New Suite of Psycholinguistic Diagnostics For Language Models. *Transactions of the Association for Computational Linguistics, 8*, 34–48.

Floridi, L. (2023). "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, & Other Generative Models". In: *Philosophy & Technology, 36*(1), 15. ISSN: 2210-5433, 2210-5441. https://

doi.org/10.1007/s13347-023-00621-y. https://link.springer.com/10.1007/s13347-023-00621-y (visited on 11/24/2023)
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines, 30*, 681–694.
Freud, S. (1921). *Die Traumdeutung*. F. Deuticke
Fuchs, T. (2020). *Verteidigung Des Menschen: Grundfragen Einer Verkörperten Anthropologie*. Suhrkamp Verlag
García-Valdecasas, M. (2022). "On the Naturalisation of Teleology: Self-Organisation, Autopoiesis and Teleodynamics". In: *Adaptive Behavior* 30.2, pp. 103–117. ISSN: 1059-7123, 1741-2633. https://doi.org/10.1177/1059712321991890. http://journals.sagepub.com/doi/10.1177/1059712321991890 (visited on 05/06/2023)
Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research, 61*, 65–170.
Glock, H.-J. (2019). Agency, Intelligence and Reasons in Animals. *Philosophy, 94*, 1–27.
Goldberg, Y. (2017). "Neural Network Methods for Natural Language Processing". In: *Synthesis Lectures on Human Language Technologies, 10*(1), 1–309
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, Massachusetts/ London, England: MIT Press.
Green, M. & Michel, J. G. (2022). "What Might Machines Mean?" In: *Minds and Machines, 32*(2), 323–338. ISSN: 0924-6495, 1572-8641. https://doi.org/10.1007/s11023-022-09589-8. https://link.springer.com/10.1007/s11023-022-09589-8 (visited on 05/06/2023)
Gubelmann, R. (2023). A Loosely Wittgensteinian Conception of the Linguistic Understanding of Large Language Models like Bert, Gpt-3, and Chatgpt. *Grazer Philosophische Studien, 99*(4), 485–523. https://doi.org/10.1163/18756735-00000182
Gubelmann, R. & Handschuh, S. (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. https://doi.org/10.18653/v1/2022.acl-long.315
Gubelmann, R., & Hongler, P., et al. (2022). "On What It Means to Pay Your Fair Share: Towards Automatically Mapping Different Conceptions of Tax Justice in Legal Research Literature". In: Proceedings of the Natural Legal Language Processing Workshop 2022. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 12–30. https://doi.org/10.18653/v1/2022.nllp-1.2. https://aclanthology.org/2022.nllp-1.2
Gubelmann, R., & Katis, I., et al. (2023). "Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks". In: *Journal of Logic, Language and Information*. ISSN: 0925-8531, 1572-9583. https://doi.org/10.1007/s10849-023-09410-4. https://link.springer.com/10.1007/s10849-023-09410-4 (visited on 12/27/2023)
Gubelmann, R., Niklaus, C., & Handschuh, S. (2022). "A Philosophically-Informed Contribution to the Generalization Problem of Neural Natural Language Inference: Shallow Heuristics, Bias, and the Varieties of Inference". In: *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*. Galway, Ireland: Association for Computational Linguistics, pp. 38–50. https://aclanthology.org/2022.naloma-1.5
Gubelmann, R., & Toscano, M. (2022). "Mechanism Versus Organism: A Loosely Kantian Perspective and Its Implications for Bioengineering". In: *Thinking: Bioengineering of Science and Art*. Ed. by N. Rezaei & A. Saghazadeh. Cham: Springer International Publishing, pp. 381–396. ISBN: 978-3-031-04075-7. https://doi.org/10.1007/978-3-031-04075-7_18. https://doi.org/10.1007/978-3-031-04075-7_18
Hassan, H., et al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: arXiv: 1803.05567.
He, P., et al. (2020). "Deberta: Decoding-Enhanced Bert with Disentangled Attention". In: International Conference on Learning Representations
Heidegger, M. (1985). "Unterwegs Zur Sprache". In: Frankfurt am Main: Vittorio Klostermann, p. 150
Jacob, P. (2023). "Intentionality". In: The Stanford Encyclopedia of Philosophy. Ed. by E. N. Zalta & U. Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/intentionality/
Jakobson, R. (2003). The Metaphoric and Metonymic Poles. *Metaphor and metonymy in comparison and contrast, 20*, 41–47.
Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review, 93*(5), 1449–1475.

Kant, I. (2012). Critique of Judgement. Trans. by J. H. Bernard. Mineola, N.Y.: Dover Publications. ISBN: 978-1-306-33778-6
Kant, I. (1785). *Grundlegung zur Metaphysik der Sitten*. Hamburg: Meiner.
Kant, I. (1781). Kritik der reinen Vernunft. Ed. by Jens Timmermann. Hamburg: Meiner
Kant, I. (1793). *Kritik der Urteilskraft*. Frankfurt a. M.: Suhrkamp.
Kassner, N. & Schütze, H. (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. https://doi.org/10.18653/v1/2020.acl-main.698. https://aclanthology.org/2020.acl-main.698
Khalid, S. (2019). BERT Explained: A Complete Guide with Theory and Tutorial. https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/ (visited on 05/22/2020)
Lan, Z. et al. (2019). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations'. In: *International Conference on Learning Representations*
Läubli, S., Sennrich, R., & Volk, M. (2018). "Has Machine Translation Achieved Human Parity? A Case For Document-Level Evaluation". arXiv preprint arXiv:1808.07048
Lavin, D. (2015). Action as a form of temporal unity: on Anscombe's Intention. *Canadian Journal of Philosophy, 45*(5–6), 609–629.
List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology, 34*(4), 1213–1242. ISSN: 2210-5433, 2210-5441. https://doi.org/10.1007/s13347-021-00454-7. https://link.springer.com/10.1007/s13347-021-00454-7 (visited on 11/24/2023)
Liu, H., et al. (2023). *Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4*. arXiv: 2304.03439 [cs]. (visited on 08/30/2023). preprint
Liu, Y., et al. (2019). "Roberta: A Robustly Optimized Bert Pretraining Approach". arXiv preprint arXiv:1907.11692
Madhyastha, P. & Jain, R. (2019). "On Model Stability as a Function of Random Seed". arXiv preprint arXiv:1909.10447
Marten, R. (1967). "„Selbstprädikation "bei Platon". In
OpenAI (2023). GPT-4 Technical Report. arXiv: 2303.08774 [cs.CL]
Poibeau, T. (2017). *Machine translation*. Cambridge, Massachusetts/London, England: MIT Press.
Popa, E. (2021). Human Goals Are Constitutive of Agency in Artificial Intelligence (AI). *Philosophy & Technology, 34*(4), 1731–1750. ISSN: 2210-5441. https://doi.org/10.1007/s13347-021-00483-2. (visited on 11/24/2023)
Radford, A., & Narasimhan, K., et al. (2018). "Improving Language Understanding by Generative Pre-Training". In: *Preprint. Work in Progress*
Radford, A., & Wu, J., et al. (2019). "Language Models Are Unsupervised Multitask Learners". In: *OpenAI Blog*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
Reis, W. J. (1970). Formen Der Freien Assoziation Zu Träumen. *Psyche, 24*(2), 101–115.
Sanh, V., et al. (2019). "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter". arXiv preprint arXiv:1910.01108
Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139173438
Shepherd, J. (2019). Skilled action and the double life of intention. *Philosophy and Phenomenological Research, 98*(2), 286–305.
Taylor, C. (2016). *The Language Animal*. Cambridge, Massachusetts/ London, England: Harvard University Press.
Touvron, H., et al. (2023). "Llama 2: Open Foundation and Fine-Tuned Chat Models". arXiv:2307.09288
van den Berg, H. (2014). *Kant on Proper Science Biology in the Critical Philosophy and the Opus postumum*. Springer.
Vaswani, A., et al. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems, 30*
Voita, E., et al. (2019). "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5797–5808. https://doi.org/10.18653/v1/P19-1580. https://aclanthology.org/P19-1580
Walsh, D. M. (2006). "Organisms as Natural Purposes: The Contemporary Evolutionary Perspective". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological*

and Biomedical Sciences 37.4, pp. 771–791. ISSN: 13698486. https://doi.org/10.1016/j.shpsc.2006.09.009. https://linkinghub.elsevier.com/retrieve/pii/S1369848606000768 (visited on 05/11/2023)

Wang, A., Pruksachatkun, Y., et al. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in Neural Information Processing Systems, 32*

Wang, A., & Singh, A., et al. (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. https://doi.org/10.18653/v1/W18-5446. https://www.aclweb.org/anthology/W18-5446

Warstadt, A., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392.

Westphal, K. R. (2014). Autonomy, Freedom & Embodiment: Hegel's Critique of Contemporary Biologism. *Hegel Bulletin, 35*(1), 56–83. ISSN: 2051-5367, 2051- 5375. https://doi.org/10.1017/hgl.2014.4. https://www.cambridge.org/core/product/identifier/S2051536714000043/type/journal_article (visited on 11/24/2023)

Wilks, Y. (2014). "Language and Communication". In: *The Cambridge Handbook of Artificial Intelligence*. Ed. by K. Frankish, & W. M. Ramsey, pp. 213–231

Wittgenstein, L. (2006). "Philosophische Untersuchungen". In: Werkausgabe Band 1. Frankfurt am Main: Suhrkamp

Wu, T., & Kang, Y.-S. (1997). Criminal Liability for the Actions of Subordinates-the Doctrine of Command Responsibility and Its Analogues in United States Law. *Harv. Int'l. LJ, 38*, 272.

Yang, Z., et al. (2019). Xlnet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems, 32*

Zammito, J. (2006). Teleology Then and Now: The Question of Kant's Relevance for Contemporary Controversies over Function in Biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 37*(4), 748–770. ISSN: 13698486. https://doi.org/10.1016/j.shpsc.2006.09.008. https://linkinghub.elsevier.com/retrieve/pii/S1369848606000756 (visited on 05/11/2023)

Žižek, S., et al. (2010). Unbehagen and the subject: An interview with Slavoj Žižek. *Psychoanalysis, Culture & Society, 15*, 418–428.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.