Precautions for Medical Decision Support by LLMs

Sebastian Bücker, Nico Formánek

Abstract *Large-Language-Models (LLMs) are currently discussed as potentially being general purpose problem solvers. One popular argument for this view is that they have been trained on almost all available text-based human knowledge (i.e. the internet) and that they therefore acquired the ability to "understand" and "reason" with said knowledge. Their ability to seemingly manipulate linguistic knowledge-representations in a correct manner, i.e. compatible with our own rational expectations, constitutes the wish to delegate rational tasks towards them. We argue that we should proceed with caution, taking into account fundamental limitations of LLMs which can be easily overlooked. Especially in environments in which relying on rational arguments can result in severe real-life consequences, like in the medical domain, such limitations need to actively be accounted for. Building on fundamental insights from the philosophy of language as well as presuppositions of current deep learning methodologies, we demonstrate what these limitations of LLMs in manipulating knowledge representations are and which precautions for their deployment into high-risk environments like medicine should be adopted. Throughout this article, we discuss our arguments with respect to the paradigmatic case of LLMs as medical decision support systems, implicitly suggesting that they can be translated into other high-risk domains.*

1. Introduction

1.1. LLMs & Medical Judgements

Large-Language-Models (LLMs) are currently debated very intensively, because apart from the language modelling capabilities they are taken to be general problem solvers. It is hard to name an aspect of human action or knowledge that is not, at least in part, represented in or working through language. The huge amount of linguistic data that contains this knowledge and which is used for training LLMs results in them being discussed as prospective tools for knowledge processing in

many areas of human expertise and for a variety of purposes.[1] What stands out is that LLMs are not used to merely predict syntactic continuations of text – their original optimization goal – but to solve tasks that are supposed to work with their conceptual content. LLMs are expected to *generate* answers which are coherent with the supplied conceptual content, i.e. having capacities that, broadly construed, can be described as *rational*. An LLM should not simply *recognize* that the user it is conversing with has uttered a specific speech act or explicated a certain intention. It is also expected to *generate* an answer.

This in turn generates the question whether they in principle possess the necessary capacities to handle the meanings of concepts. Do LLMs actually have some kind of conceptual "knowledge" that they can apply? Have they captured the implicit and/or explicit "'rules" that legitimize how we use language, especially how we *move* from one proposition to another? The "rules" we have in mind are not just formal or logical rules like the *modus ponens*,[2] but ones that necessitate to consider a multitude of aspects: the complexity of the situation that the speaker might be in, his previous commitments, common and expert knowledge, the situation of the audience, and so on, i.e. the whole range of concept use in language. Imagine a doctor telling a patient's diagnosis to another doctor: He is having certain *general* medical knowledge himself, which might not completely overlap with that of whom he is talking to. He has *concrete* knowledge about the patient and his conditions. He might previously have given a different diagnosis, knowing that the doctor he is talking to heard him commit himself to this. By forming his new diagnosis, i.e. forming a judgement, he accounts for all these aspects by giving his reasons for why he is retracting his previous diagnosis, saying which facts changed or which general medical knowledge he since acquired that moved him to reevaluate his concrete situation. A complex network of rules, implicit and explicit, is governing how commitments[3] are made.

It has already been shown that GPT-4 is able to answer the questions from the U.S. Medical Licensing Examination (USMLE) with an accuracy of more than 90% (Lee, Bubeck, and Petro 2023). The fact that LLMs show this kind of "expertise" has motivated research into prospective medical applications for GPT-4 and other LLMs, e.g. as medical consultants that could substitute real doctors when advising patients such that they can give their *informed consent* (Kienzle et al. 2024; Rao et al. 2025). Due to relatively easy deployment combined with widespread availability, LLMs are also discussed as prospective tools to shift towards more of a personalized and predictive approach to medicine, compared to previous ones that just aim at treating

1 Thus, falling under the definition of "General Purpose AI" of the EU AI Act (cf. Future of Life Institute 2024).

2 This rule of inference moves from the premises A and $A \to B$ to the conclusion B.

3 Herein, commitments are to be understood as linguistic propositions that one utters in order to be determinable towards other speakers.

pathologies (Nogaroli et al. 2024). But many remain skeptical as to whether LLMs have really mastered the full scope of linguistic rules that govern how we formulate commitments, since LLMs exhibit what has been called "strange errors" (Rathkopf and Heinrichs 2024) or "hallucinations" (Peng, Narayanan and Papadimitriou 2024). Such hallucinations seem to be a feature of the architecture and can contrary to earlier hopes never be ruled out completely (ibid.). There is thus a case for forward-looking responsibilities, i.e. precautions that any user of LLM-mediated decision support has to take (Sand, Durán and Jongsma 2022). Many authors motivate the necessity of such precautions by pointing to LLM outputs with occurrences of adversarials or hallucinations, because they are "raising concerns about the consistency and reliability of responses." (Kienzle et al. 2024: 8). Common to such arguments is then the call to validate the results that LLMs produce. For example, if patients interact with the LLM, then "healthcare professionals should continuously evaluate ChatGPT's responses to common patient queries, ensuring its reliability and relevance in a clinical setting." (ibid.: 8) or that "[p]hysicians need to critically assess what output values are reasonable given certain input values." (Sand, Durán and Jongsma 2022: 167).

The authors above are not explicit about what exactly a validation should entail and how it can be achieved. Thus, it is unclear which precautions a medical professional has to take. In the following sections we will argue that certain technical details of LLMs impose specific precautions on medical professionals. Practicing these precautions when conversing with a chatbot would constitute a capability to take an LLM's utterance not at face value, but knowing which aspects to *re*consider in order to make the utterance one's own and *take* responsibility for it. Such a capability would, for example, be fitting for the role of a "clinical AI Expert", as it has been proposed by Bartsch et al. (2025) in this volume.

1.2. LLMs and Chatbots

Many ethical texts do not distinguish explicitly between general decision support systems, LLMs in a classification setting and LLMs in a chatbot setting (Sand, Durán and Jongsma 2022 for example just talk about medical AI). While this might be unproblematic if one is only interested in the ethics of decision support, we are specifically worried about the inherent shortcomings of LLM chatbots in medicine. The medical domain is particularly suited for reflecting these shortcomings, as the patient's vulnerability demands a rigor of argumentation – the inductive risk is high (Karaca 2021). Its subject matter is at the same time associated with an inherent complexity, one that demands a challenging education in order to get a partial epistemic overview of. Our arguments also translate to other domains, but the contrasts we wish to make become especially visible in the medical domain. We thus would like to emphasize that the situation we have in mind when we speak about medical LLMs

is the decision support by a chatbot LLM (e.g. ChatGPT) for a medical professional. This LLM might be a foundation model finetuned on medical data or just a vanilla transformer type model.

One of the chatbot scenarios we take as a background is discussed in Saenger et al. (2024). Although the LLM usage therein is not from a medical professional, it nevertheless illustrates the dangers of missing epistemic capabilities. They report a man in his 60s, who presented himself to the emergency department (ER) with multiple episodes of diplopia (double vision). The first episode was characterized by his interventionalist as a harmless aftereffect of a prior heart surgery (pulmonary vein isolation). But the patient was not satisfied with the explanation that his physician gave him. So, he decided to consult ChatGPT (version 3.5), describing his situation and symptoms in a similar manner as he did towards the physician. Relieved that ChatGPT came to the same conclusion, while giving much better explanations, he felt content and did not pursue further actions. After all, both his physician and ChatGPT told him that his visual disturbances were harmless aftereffects, which are bound to improve shortly. In the ER, a more thorough investigation of his symptoms and medical history indicated an acute stroke. This prompted further investigations and intensive care at the local stroke unit, later resulting in an updated, less drastic diagnosis of a transient ischemic attack (sometimes called a "mini stroke"). This example not only shows a case of LLM decision support drastically gone wrong, but more importantly illustrates that a chatbot was taken as semantically on par with a physician. It came to conclusions, gave explanations and supplied diagnoses that were considered meaningful. It highlights the necessity of having undergone a demanding education to acquire the capability to really make sense of medical propositions. In the next section we will discuss in detail how one should think about the semantics of LLM-generated text.

2. Language Use and LLMs in Medical Contexts

2.1. Distributive Semantics

Perhaps one of the main hopes regarding LLMs capturing semantics, is the "distributional hypothesis", which states that a word's meaning can be characterized by its linguistic context in which it usually appears (Grindrod 2024). The intuition is that the frequency distribution of a word's contexts already contains enough information to implicitly fix its meaning(s). Obviously, this rather vague idea needs to be operationalized in code. The usual approach, also employed for current LLMs, is to represent words or more correctly tokens – a smaller linguistic unit – as vectors in a high-dimensional space. The specific vectors encoding a token are calculated during the training run of the LLM. They are thus only calculated once during training

and do not change during inference. LLMs are trained by optimizing a specific task, most often slightly different variations of the masked-language prediction task. The optimization goal for this task is purely syntactical, namely, to best predict masked sections of a given text. Any "meaning" that is encoded into the vector representation of a token is thus at best implicit and often does not correspond to what humans consider meaningful (Church 2017).

The methodology behind LLMs does not apply the idea of distributional semantics in any principled way. Engineering considerations have arguably influenced the design of the transformer architecture much more than theoretical ideas about semantics. This also means that the pragmatic success of LLMs is at best indirect evidence for the distributional hypothesis. One might circumnavigate the issue and just say because vector representations "provide the basis for the state-of-the-art across a whole host of other meaning-related tasks" (Grindrod 2024: 71), they in fact do work as representations for meanings. In other words: The distributional properties captured in the vector representations and learned weights of the neural components of an LLM can, in many cases, serve as substitutes for the semantic properties of the represented words. But obviously, they only serve as such as long as we accept the generated text as meaningful. But as long as we cannot guarantee a priori that generated text will be meaningful, we have to rely on either our immediate evaluation or (standardized) benchmarks. Such benchmarks show a whole host of things but are crucially at best proxies for meaning (Mitchell and Krakauer 2023). Besides the aforementioned attempts at finding empirical evidence of meaning in generated texts, the distributional hypothesis still seems to be the best shot at giving a theoretical explanation of how it might enter generated texts (Mollo and Millière 2023). This is why we discuss a theoretical argument against the distributional hypothesis in the next section.

2.2. A Wittgensteinian Argument against Distributive Semantics

The approach to language that the latter Wittgenstein proposed can superficially be viewed to be in accord with distributive semantics. He writes in the *Philosophical Investigations*: "the meaning of a word is its use in the language." (Wittgenstein 2010: §43) This has been construed such that "use in language" was to mean something analogous to the distributional hypothesis (cf. Lenci 2008). And the Firthean slogan "You shall know a word by the company it keeps!" (Firth 1962: 11) seems to suggest the correctness of the analogy. But such an interpretation ignores Wittgenstein's remarks on how meaning is connected to what he calls *form of life* (*Lebensform*). He famously wrote: "What is true or false is what human beings say; and it is in their language that human beings agree. This is agreement not in opinions, but rather in form of life." (Wittgenstein 2010: §241) His concept of *Life Forms* serves to emphasize how entangled our linguistic and non-linguistic activities are in our cultural activity

of using a language (cf. Schulte 1992: 110 et seq.).[4] One could think of "how a clarinet sounds" (Wittgenstein 2010: §78) and, for example, linguistically *describe* their sound similar to that of a laughing or crying person. Wittgenstein observes that this merely shifts the interpretation of what a clarinet sounds like to *another* linguistic expression. If, for example, a child never heard the sound of a clarinet, and its parents were to give the answer above, the child needs to already know what a laughing or crying person sounds like. And now the same question reappears: Would it be possible to *describe* this linguistically? At some point, such linguistic explanations would need to stop and *something* categorically different needs to be given, or else the interpretations would somewhere start to run in circles. For Wittgenstein, this "gap" can be only filled by non-linguistic, *practical knowledge*: One *knows* what a clarinet sounds like, if one has heard one. And this does not imply the capability to also *describe* such a sound (beyond being one of a clarinet).[5]

Now, since LLMs *only* have access to the distributional properties of language, i.e. statistical properties of an enormous corpus, they can by design not incorporate the non-linguistical activities that are otherwise entangled with our language use. Wittgenstein's paradigmatic situation, in which such entanglement shows up, are pain ascriptions, which he analyses intensively in his *Philosophical Investigations*. His remarks target the difficulties and asymmetries between first-personal and third-personal utterances. Saying "I am in pain" has a different meaning or function, i.e. it is a different *move* in the language-game than "He/She/It is in pain". An "obvious" reason for this could be that from a first-personal perspective one has epistemic access to one's inner feelings, from which any third personal speech is epistemically excluded. But Wittgenstein resists this interpretation with his *private language arguments*. It cannot make *sense* that such utterances are *directed towards* inner episodes, because language is essentially a social enterprise. Somebody hearing such an utterance would *necessarily* need non-linguistic clues serving as a criterion for the identity of what is represented and its representation.

Rather than understanding pain ascriptions as representing inner episodes, Wittgenstein emphasizes how the usage of such utterances needs to have *practical* consequences: "And now look at a wriggling fly, and at once these difficulties vanish, and pain seems able to get a foothold here, where before everything was, so to speak, too smooth for it." (Wittgenstein 2010: §284) If someone has pain, this

4 Grindrod (2024: 71) also notes in a footnote that "Wittgenstein envisioned use as understood both in terms of linguistic and non-linguistic activities". Given his ambition to "consider whether LLMs meet the conditions prescribed by our best metasemantic theory", it seems as though he commits a *petitio principii* by not considering those metasemantic theories that do incorporate non-linguistic activities. It seems like his argument already presupposes the correctness of the distributional hypothesis implicitly.

5 Similar observations are due to Ryle (2009: Ch. 2), who introduced the distinction between *knowledge-that* and *knowledge-how*. For a comprehensive overview, see Brandom (1994: Ch. 1).

typically also shows in their behavior. Children that are yet to learn how to speak would, for example, cry, thereby drawing the attention of their parents towards them, moving them to alleviate their pain. Wittgenstein (ibid.: §244) asks, how does it come about that in such situations, a child learns how "words *refer* to sensations". A *possible* answer is that words, in our current example the names of certain kinds of sensations, get used instead of the respective behaviors. The child learns that saying a certain pain name generates the same responses in their parents as does their pain behavior. "[The] verbal expression of pain replaces crying, it does not describe it" (ibid.: §244). This sketches Wittgenstein's position on what it means to use names of sensations from a *first-personal perspective*. However, using names of sensations from a *third-personal perspective*, i.e. by ascribing *to* somebody that he or she has a certain sensation, has a different practical meaning. Such an ascription needs to be connected to a range of other commitments that accompany the ascription. These can at first be implicit and only be called upon to show incompatibilities. Ascribing pain to a dead person would create such an incompatibility, because dead persons do not exhibit pain behavior, just as pain is not ascribed to the respective body part in which the pain "resides". Only the "whole" person exhibits such pain behavior, either by speaking or by acting in certain ways.[6] By making an ascription, one needs such accompanying commitments that serve to identify the respective person. Regarding one's own pain or other sensations, all such criteria that serve to identify a person *are* ignored. I simply *have* them. They might differ with respect to certain other qualifications, like intensity or in which body part they occur, but it is always the *undifferentiated* "I" who has them.

For Wittgenstein, such kinds of pragmatic[7] embedding of our language use with non-linguistic aspects are the "bedrock" for what our utterances *mean* (ibid.: §217). Wittgenstein stresses that this is, in a certain sense, opposed to an understanding in which names of sensations *refer* to inner episodes. This does not mean that inner episodes do not exist, but that it is meaningless to want to *directly* speak about them.[8] In contrast to worldly things, one could not point towards them using the determiner "this", since, as Wittgenstein puts it, "one does not define a criterion of identity by emphatically enunciating the word 'this'." (ibid.: §253) Another person would not be able to identify "this" pain, if no other pragmatic hints are given, i.e.

6 For Wittgenstein, the non-linguistic, pragmatic aspects of language use are constitutively linked to specifically *human* forms of action: "If a lion could talk, we wouldn't be able to understand it." (Wittgenstein 2010: Part II *A Fragment xi*, §327)

7 Understood here loosely as in the tradition of *pragmatism*: "[T]he core is the belief that the meaning of a doctrine
is best understood through the practices of which it is a part." (Blackburn 2016: 374 et seq.)

8 "[Sensations are, S.B.] not a Something, but not a Nothing either!" (Wittgenstein 2010: §304)

that the person does not exhibit non-linguistic pain behavior or linguistically defines the pain. It is only due to entangled connections of words *with pragmatic consequences*, e.g. our own and other's actions and behavior *while* speaking, that a word *has* meaning. Wittgenstein's position is permeated by an anti-Cartesian stance that rejects the priority of inner experiences to explain the meanings of words, instead they are formed by our social, interpersonal, and cultural *Life Forms*. If one accepts Wittgenstein's analysis, this points towards aspects of language meanings that can neither be learned nor generated by LLMs. Specifically pain ascriptions, which can play a central role in medical diagnoses, are affected by this. Disregarding the non-linguistic aspects, as it is done in distributional semantics, does not give a feasible way towards grounding the meaning of language. It would, in Wittgenstein's words, "hang in the air" (Wittgenstein 2010: §198) without support, but without it, medical LLMs are disconnected from the life forms of medical practitioners.

The next section illustrates that the preceding philosophical considerations have direct consequences in LLM-supported medical decisions. It shows that attempts to validate such decisions have to take into account the *pragmatics* of language (the life forms) and cannot be conducted at the linguistic level alone.

2.3. De Re and De Dicto Aspects in Language Use

Our language use can often be a source of ambiguities that arise due to the possibility of different contexts in which an utterance can be evaluated. Luckily, very often our languages themselves offer the means to mitigate such ambiguities. A possible source of ambiguities which often arises in speech is that it can be unclear how to distinguish between what is being said, i.e. the predicative function of a proposition, and which thing, pointed towards by the grammatical subject, this predicative function says something about. These locutions can occur conflated, because it often is necessary to describe the object (grammatical subject) of a proposition using words that can also have a predicative function, e.g. "the blue sky has no clouds" in which "blue" is not part of the predicative function that is expressed, but part of the term that denotes the subject of the predicate. This gives rise to the distinction between *de re* and *de dicto* aspects of propositions,[9] due to which it can have different interpretations.

The philosophical interest in these two aspects stems from the fact that when they appear conflated, it becomes difficult to substitute different identifiers of one and the same object. Quine (1953) gives the example of the Italian painter Giorgione, whose name in the proposition "Giorgione was so called because of his size" could

9 These latin phrases translate to "of a thing" (*de re*) and "of a statement" (*de dicto*), see Blackburn (2016: 128).

not be substituted with his middle name, Barbarelli, without making the proposition false. Because Giorgione is the augmentative of Giorgio, i.e. a morphological form of the Italian language that makes the thing it denotes bigger,[10] the truth value of the proposition stems from the fact that precisely the name Giorgione is used to denote the grammatical subject. This creates a problem since, at least in principle, when something is true about an object, this should not depend on whether we choose a different identifier to denote the object. A conflated use of *de re* and *de dicto* in a proposition can violate the principle of truth conservation under substitution, which plays a fundamental role in logic. But such a conflated use can, in the example above, be disambiguated via substitution of the anaphoric "so" by the name in single quotes, thereby making the initial identifier Giorgione substitutable again: "Barbarelli was called 'Giorgione' because of his size",[11] in which the term 'so' has been substituted by 'Giorgione'. In this way, the ambiguity can be mitigated via another linguistic explication of the same idea.

The example given above provides an ambiguity of *de re* and *de dicto* aspects even if there is a singular context of interpretation. This means no controversial arguments exist between speakers with respect to the subsentential elements, like words and phrases, that are used. But such a singular context can, in most cases, not be assumed, especially when different persons with a difference in perspective, knowledge, and sets of beliefs interact. Linguistically, such contexts are marked using phrases that ascribe a propositional attitude to a person, as in the paradigmatic form "she/he beliefs that …". This expresses that from the respective person's perspective, i.e. the one to whom a proposition is ascribed, something can be described 'so-and-so'. But this might not necessarily be the case for another person's perspective, especially the one who ascribes the belief. Think of a doctor who tells his or her patient "I believe you are having a transient ischemic attack". This means something different to the doctor in a number of aspects: He (hopefully) knows how to cure or alleviate the attack, he knows for which other diagnoses such an attack can be a symptom and by which methods these could be investigated. The patient, however, might not even know what a transient ischemic attack is, prompting the doctor to explain it to him, to which the doctor could reply that it is a "mini stroke". Such differences in perspective, knowledge and belief need to be taken into account when validating propositions that express other propositional attitudes, like "she/he beliefs that …".

10 Another example: the augmentative of *porta* (door) would be *portone* (gate).

11 This sentence is the result of two substitutions: Starting from "Giorgione was so called because of his size", substitute the anaphoric "so" with "'Giorgione'" in single quotes. This results in "Giorgione was called 'Giorgione' because of his size", which now allows to substitute the mention of *Giorgione* without single quotes by *Barbarelli*.

The capacity to differentiate between such kinds of perspectives, in medical contexts, roughly the patient's perspective and the doctor's perspective, would be vital for LLMs. It should 'know' how to address its interlocutor, e.g. whether it can communicate with him or her using idiosyncratic medical terms or not. Situations in which an LLM would need to handle such differences in perspective will easily occur. A use case in which LLMs are substituting real physicians that is currently researched, is informed consent. Prior to participation in a medical study, patients partake in a conversation with their physician in order to ask questions and to orient themselves on which effect the intervention might have on their life as a whole, not just in a medical sense.

The patient's reasons for consent would likely mention propositional attitudes (e.g. beliefs, knowledge) of their physician, or those of other physicians from the past, that are not involved in the current treatment. An LLM that should substitute the respective physician in such a conversation would need the capability to disentangle all these different perspectives when conversing with the patient. Such situations can also occur in LLMs that are supposed to communicate with medical practitioners. The LLM might be primed, e.g. by prompt engineering, that it is conversing with a medical practitioner. Suppose the practitioner told the LLM that "the patient believes that he has a transient ischemic attack", which could make sense, even if the patient did not know what having a transient ischemic attack means. But the LLM cannot disentangle the *de re* and *de dicto* aspects of the sentence. It cannot distinguish between the doctor's ascribed diagnosis and the patient's perspective. Brandom (1994: Ch. 8) argues, that the *de re* and *de dicto* aspects can disambiguate such propositions explicitly. What creates the ambiguity in the proposition is that the practitioner ascribes his own interpretation of the patient's symptoms to the patient. One might be misled that the patient has commitments with respect to this proposition, for example that he could give reasons for the belief that his symptoms can be characterized as a transient ischemic attack. But, in this constructed scenario, the patient does not know what such an attack is, much less how to identify one. To him, his symptoms constitute a mild stroke. A disambiguated expression – uttered by the practitioner – should be "the patient believes of his transient ischemic attack, that it is a mild version of a stroke". This, according to Brandom (ibid.), would make it clear that the practitioner is responsible for the interpretation as transient ischemic attack, while the patient has a commitment to explain why he thinks that his symptoms are those of a stroke. Using *de re* and *de dicto* aspects in such a disambiguated form can serve the purpose of making clear who is responsible for which interpretation.

Connecting this to the Wittgensteinian arguments presented above, two aspects of the ability to validate propositions in medical contexts can be discerned: First, the ability to distinguish the *de re* and *de dicto* aspects of a proposition. This includes both a capability to correctly disambiguate the respective predicative function and

grammatical subject. As argued above, this means that different perspectives are necessary for a correct interpretation, requiring knowledge of who is responsible for interpreting in the first place. This capability is closely related to what currently is being researched as Theory of Mind capabilities of LLMs, i.e. whether they can reason "about other people's intentions, goals, thoughts, and beliefs" (Sclar et al. 2024: 1). This could, for example, be false belief tests like Sally-Anne-Tests, in which the models would need to reason from the perspectives of others, which often might not correspond to actual facts. Most LLMs exhibit good capabilities in fairly easy situations, but Sclar et al. (ibid.) showed that in more complex situations, the performance of current models drops significantly, with models like Llama and GPT-4o achieving accuracy values of below 10%.

Secondly, as Wittgenstein's arguments for a pragmatic embedding showed, correct interpretation also includes an ability to incorporate non-linguistic aspects. After knowing who needs to be approached for an interpretation, one also needs to know which pragmatic truth makers can be called upon and how these need to be distributed between different perspectives. An LLM could generate text stating that a transient ischemic attack might show itself with symptoms of double vision, a weakness up to paralysis in one side of the body's limbs or impairments of speech. If an LLM suggests checking for one of those, given that the others have already occurred, this means that descriptions of transient ischemic attacks significantly often mention these symptoms together. But it would still require a physician who, after hearing that the patient has double vision, would touch the patient's limbs, while also asking the patient how that feels. This shows how entangled linguistic and non-linguistic activities are when validating medical judgments.

3. Updating Beliefs Responsibly

3.1. Brandom's Discursive Responsibility Dimensions

Brandom, in the tradition of Wittgenstein, further examines the entanglement of linguistic and non-linguistic activities, especially with regard to their *social aspects*. A crucial point in Brandom's thought is that he does not think about concepts from the perspective of how we can *know* and *apply* them, but rather which implicit commitments our use of them implies. Thus, his primary object of investigation is deeply entangled with how the use of concepts makes the speaker *responsible* for what he said. Brandom's philosophy is deeply rooted in a change of perspective on what it means to be conscious. Classical positions hold that consciousness is directed towards its object and having *good* conceptual representations means that they can be *successfully* used to anticipate states of affairs. Brandom adds to that a direction *from*

the object *towards* the subject, since the object "exercises a special sort of *authority*" (Brandom 2009: 34) upon the subject.[12]

This aspect is important for interpersonal, i.e. social language use, in which each participant uses concepts and thus makes himself responsible towards the other for what is said. For Brandom, the concept of a *belief*, i.e. a proposition one takes to be true, is to be understood as an *inferential commitment*: It is inferential in the sense that it allows others to query the speaker with conceptually informed questions regarding his reasons or conclusions, and it is a commitment since speaking is conceptually an act by which the speaker enters into certain obligations. The interlocutor becomes the "object" that exercises authority upon the speaker.[13] Brandom identifies three distinct basic forms of such discursive responsibilities. The first is the speaker's "*critical* responsibility to weed out materially incompatible commitments." (ibid.: 36) In short, this is the *law of non-contradiction* for material inferences (i.e. linguistic inferences that are warranted *not* by formal logic). If a doctor *infers* from "this mole is malignant" that "the mole needs to be excised", this inference is not warranted by his knowledge of a conditional like "If a mole is malignant, then it needs to be excised" and applying *modus ponens*. The doctor can *explicate* his knowledge by such a conditional, but the conditional is not what warrants the inference. Thus, inferential reasoning is not primarily a logical capability, and non-logical aspects play a major role in it. A speaker has the central responsibility of "aiming at a whole constellation of commitments that is *consistent*." (ibid.: 36)

The next dimension of discursive responsibility is the "*ampliative*", which means that a linguistic commitment entails other commitments, to which the speaker implicitly committed himself in the first place. It is his responsibility to extract such *material* consequences and to not deny his commitment (or argue against the entailment). Fulfilling this responsibility then aims at a constellation of commitments that is *complete* (ibid.: 36). The last responsibility reverses this perspective towards the *reasons* for one's commitment: "One's *justificatory* responsibility is to be prepared to offer reasons for the commitments (both theoretical and practical) that one acknowledges" (Brandom 2009: 36), aiming at a *warranted* constellation of commitments.

The case reported by Saenger et al. (2024; discussed in section 1.1) can serve to illustrate how these responsibilities interact in order to get to an updated constellation of commitments. The patient described his situation and symptoms to ChatGPT, which told him that "in most cases, visual disturbances after catheter ablation are temporary and will improve on their own within a short period of time" (ibid.:

12 Brandom does not think of himself as being the first to investigate such a theory of concepts and calls on Kant, Hegel, Frege and Wittgenstein as earlier exponents (cf. Brandom 2009).

13 Think, for example, of how giving a promise means that its receiver *can* call on you to fulfill your commitment.

237), thereby calming the patient. However, as Saenger et al. (ibid.) note, prompting ChatGPT whether visual impairments after catheter ablation (the prior operation of the patient) could indicate a stroke, ChatGPT affirmed this. This constitutes what Brandom calls a material incompatible constellation of commitments: One cannot *act* both linguistically and non-linguistically according to such commitments. Telling somebody that he might have a stroke, entails that he should seek immediate medical attention. According to Brandom, having such an incompatible constellation of commitments *should* move the respective speaker to change his or her constellation of commitments. In this way, the *ampliative* dimension was used to show how there was an incompatible constellation of commitments. In order to update one's constellation, it does not suffice to just withdraw from one commitment that has been shown to create the incompatibility. Here, the *justificatory* dimension comes into play: One needs good reasons that inferentially warrant and explain the new constellation of commitments. In the case that Saenger et al. (ibid.) describe, this happens since the physicians that treated the patient describe the methods to confirm the diagnosis of his condition (MRI scans, computer tomography, etc.). *This* allowed them to *change* their commitment: "Therefore, the working diagnosis was changed to TIA" (ibid.: 237). This change of one's constellation of commitments can again be made explicit using the *de re* and *de dicto* aspects, discussed above: 'At first, the physicians believed *of* a transient ischemic attack, *that* it was an acute stroke.' The physicians *reasonably* changed their perspective, they changed the constellation of commitments towards the patient. This constitutes a *linguistic act* that we believe LLMs, for technical reasons, cannot carry out. We will present an argument for why LLMs cannot perform such acts in the next section.

3.2. The Inability of LLMs to Change their Beliefs

Many philosophers have wondered if LLMs have internal states and can update these internal states. These discussions are obviously related to questions about beliefs, belief revision, and other intentional states (meaning, knowledge, commitments) of LLMs. The following argument will be mainly concerned with beliefs, because this is the intentional state that has been mostly discussed in philosophical considerations of LLMs. We believe that it applies equally to other intentional states.

Because it is central for the following argument about "belief-change" in LLMs, we already note here that we only consider the standard GPT architecture (for example as elucidated by Phuong and Hutter 2022). Specifically, this means that no external memory is supplied and there are no wrappers implementing techniques like retrieval augmented generation[14] – having such wrappers or access to memory

14 It is not completely clear to us if the currently popular reasoning models like OpenAI's o3 or DeepSeeks R1 break our architectural assumptions. From what is known publicly it seems that

would give the LLM the possibility to save states across inferential steps, effectively constituting an internal state.

It is very helpful to separate the question if LLMs have beliefs from the question if they can change their beliefs. Because the answer to the second one is a clear no. The answer to the first question is less clear, but because LLMs – during inference – are static functions, the following implication holds: If you think that a static object like a book can hold beliefs, you must also say that an LLM can hold beliefs. As always, one man's *modus ponens* is another man's *modus tollens* – if you deny that LLMs can hold beliefs out of hand, then you must also deny that books can hold them. We will thus bracket the question if LLMs can hold beliefs and focus on the question whether LLMs can change their beliefs. While the interaction with possibly belief-holding entities like books has been considered unproblematic in medical diagnostics – think about manuals for differential diagnosis – the usage of LLMs has come under close scrutiny. This is not at all bad. LLMs give the impression, especially if used in interactive chatbot settings, that their outputs are actually the locutions of an agent capable of tracking and updating the conversation with an internal state. At least they give this impression sometimes (Saenger et al. 2024), even though there is also ample evidence to the contrary (Hager et al. 2024). This impression is problematic, we claim, precisely because we expect behavior of a chatbot similar to that of a person, capable of revising their internal (i.e. mental) state(s). If we pointed out their mistakes to someone who cannot, by their nature, change, we cannot blame them – it does not even make sense to speak of Brandomian commitments. But most importantly we cannot *expect* them to change. We thus cannot expect LLM-based chatbots to change during the conversation. This will hold for every property that requires the update of an internal state, be it knowledge, belief, meaning, representation, or commitments. This means LLM-based chatbots also cannot "learn" anything during a conversation. Now one might think, this is not how these systems have been discussed in the media. Wasn't the upshot of modern AI that these systems adapt and are actually capable of learning? To answer this question some details on the specific technology we are discussing are necessary. There are two stages which are important in the life of an LLM. They are trained and they are used for inference. The important thing to notice is that, in case of current LLMs, these two stages are completely separated in time and one of them, the training, is never repeated. After it is completed, an odd trillion of internal weights have settled and what we have is nothing more than a very complicated mathematical expression which we can use for inference. During inference these internal weights do not change ever again.[15] What

only the training process is different, while the inference still follows the Markov dynamics of the basic GPT architecture.

15 This observation can also be stated in a very technical way by noting that LLMs during inference are finite state, finite order Markov chains (Zekri et al. 2025).

changes is only the input to the mathematical expression, the prompt. So, whenever an LLM-based chatbot gives the impression of learning, of revising its beliefs or of gaining knowledge, it is just a function of the prompt and random sampling. But neither the function nor the random sampling ever changes.

Thus, if one thinks the ability to change one's internal state(s), to "update" one's beliefs, is important for giving good advice and being responsible for giving bad advice, then chatbots, as they are constructed now, lack exactly this ability. They are thus, for technical reasons, unable to perform that linguistic act that concludes in changing their commitments. An informed user must be aware of this fact. A precaution directly following for medical professionals from this technical property is that they constantly need to be aware of that they are interacting with a system which can neither change its beliefs nor its commitments.

4. Conclusion

The aforementioned arguments can be summarized into the following view about what LLMs essentially are: They are a *symbolic inferential picture* of our linguistic practice. They are symbolic since they work only on symbols and their combinations. They are inferential[16] since they transition between a finite set of states according to a predefined rule. They are pictures because their architecture is *static* and cannot change during conversations.

Incorporating these aspects into medical decision making requires professionals to cultivate a precautionary stance towards the outputs of LLMs. Their utterances should not be taken at face value. This might sound more trivial than it is. In doing so, the inherent inabilities of LLMs have to be *actively* accounted for – especially when their results get *incorporated* into one's own commitments. This requires constant vigilance.

Thus, a responsible integration of LLMs into medical practice needs to take the following specific precautions. Medical practitioners need to *disambiguate* the *de re* and *de dicto* aspects of outputs generated by an LLM. Wittgenstein's discussion of pain ascription showed how this necessarily involves non-linguistic aspects of the medical "life form". Therefore, LLM outputs need to be validated in practice. But this practice must not be mistaken as a practice where one converses with a counterpart who has the ability to change their beliefs. The precaution that has to be taken against this view consists in understanding the technical details of LLMs at such

16 This usage of "inferential" in the sense that what they are designed to and seem to do is inference. This should not be confused with the notion of "inferentialism" that Brandom advocates. Our point is precisely that LLMs do not "infer" in the strong sense that Brandom explicates, i.e. inferential *reasoning* (cf. Brandom 1994; Brandom 2001; Brandom 2009).

a level that their inability to change beliefs becomes evident. Medical institutions should facilitate the cultivation and integration of such a precautious stance, whenever an LLM is used in a decision support role. This could for example be enabled by a governance structure, akin to the one proposed by Bartsch et al. (2025) in this volume, where a "clinical AI expert" mediates between the technical and medical forms of life.

References

Bartsch, Sebastian et al. (2025): "Ethics and Regulation of AI Systems in Medicine. The Example of Cancer Detection", in: Kaminski, Andreas et a. (eds.), Trust and Responsibility. Digital Governance from a Capability-Oriented Perspective, Bielefeld: Transcript.

Blackburn, Simon (2016): The Oxford dictionary of philosophy, Oxford: Oxford University Press.

Brandom, Robert B. (1994): Making It Explicit. Reasoning, Representing, and Discursive Commitment, Cambridge, MA: Harvard University Press.

Brandom, Robert B. (2001): Articulating Reasons. An Introduction to Inferentialism. Cambridge, MA: Harvard University Press.

Brandom, Robert B. (2009): Reason in Philosophy. Animating Ideas. Cambridge, MA: Belknap Press of Harvard University Press.

Church, Kenneth W. (2017): "Word2Vec", in: Natural Language Engineering 23, pp. 155–162.

Firth, J.R. (1962): Studies in Linguistic Analysis. Oxford: Basil Blackwell.

Future of Life Institute (2024): "High-level summary of the AI Act", https://artificialintelligenceact.eu/high-level-summary/, last access: February 02, 2025.

Grindrod, Jumbly (2024): "Large Language Models and Linguistic Intentionality", in: Synthese 204.

Hager, Paul et al. (2024): "Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making", in: Nature Medicine 30, pp. 2613–2622.

Karaca, Koray (2021): "Values and Inductive Risk in Machine Learning Modelling. The Case of Binary Classification Models", in: European Journal for Philosophy of Science 11, p. 102.

Kienzle, Arne et al. (2024): "ChatGPT May Offer an Adequate Substitute for Informed Consent to Patients Prior to Total Knee Arthroplasty – Yet Caution Is Needed", in: Journal of Personalized Medicine 14, p. 69.

Lenci, Alessandro (2008): "Distributional Semantics in Linguistic and Cognitive Research", in: Italian Journal of Linguistics 20.

Mitchell, Melanie and David C. Krakauer (2023): "The Debate Over Understanding in AI's Large Language Models", in: Proceedings of the National Academy of Sciences 120. arXiv: 2210.13966 [cs].

Mollo, Dimitri C. and Millière, Raphaël (2023): "The Vector Grounding Problem" arXiv: 2304.01481 [cs].

Nogaroli, Rafaella et al. (2024): "Ethical Challenges of Artificial Intelligence in Medicine and the Triple Semantic Dimensions of Algorithmic Opacity with Its Repercussions to Patient Consent and Medical Liability", in: Sousa, Henrique A. et al. (eds.), Multidisciplinary Perspectives on Artificial Intelligence and the Law, Cham: Springer International, pp. 229– 248.

Peng, Binghui, Srini, Narayanan and Papadimitriou, Christos (2024): "On Limitations of the Transformer Architecture", arXiv: 2402.08164 [stat].

Lee, Peter, Bubeck, Sebastien and Petro, Joseph (2023): "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine", in: The New England Journal of Medicine 388.Phuong, Mary and Hutter, Marcus (2022): "Formal Algorithms for Transformers", arXiv: 2207.09238 [cs].

Quine, Willard Van Orman (1953): "Reference and modality", in: From a Logical Point of View. Cambridge, MA: Harvard University Press, pp. 139–159.

Rao, Vishwanatha M. et al. (2025): "Multimodal Generative AI for Medical Image Interpretation", in: Nature 639, pp. 888–896.

Rathkopf, Charles and Heinrichs, Bert (2024): "Learning to Live with Strange Error. Beyond Trustworthiness in Artificial Intelligence Ethics", in: Cambridge Quarterly of Healthcare Ethics 33, pp. 333–345.

Ryle, Gilbert (2009): The concept of mind, ed. by Tanney, Julia, London: Routledge.

Saenger, Jonathan A. et al. (2024): "Delayed Diagnosis of a Transient Ischemic Attack Caused By ChatGPT", in: Wiener klinische Wochenschrift 136, pp. 236–238.

Sand, Martin, Durán, Juan M. and Jongsma, Karin R. (2022): "Responsibility Beyond Design. Physicians' Requirements for Ethical Medical AI", in: Bioethics 36, pp. 162–169.

Schulte, Joachim (1992): Wittgenstein: An Introduction. Albany, NY: SUNY Press.

Sclar, Melanie et al. (2024): "Explore Theory of Mind. Program-guided Adversarial Data Generation For Theory of Mind Reasoning. arXiv: 2412.12175 [cs].

Wittgenstein, Ludwig (2010): Philosophische Untersuchungen: = Philosophical investigations, trans. by G. E. M. Anscombe, P. M. S. Hacker and Schulte, Joachim, Chichester: Wiley-Blackwell.

Zekri, Oussama et al. (2025): Large Language Models as Markov Chains. arXiv: 2410.02724 [stat].