Rights and Wrongs in Talk of Mind-Reading Technology

Rainey, Stephen

DOI
10.1017/S0963180124000045

Publication date
2024

Document Version
Final published version

Published in
Cambridge Quarterly of Healthcare Ethics

Citation (APA)
Rainey, S. (2024). Rights and Wrongs in Talk of Mind-Reading Technology. *Cambridge Quarterly of Healthcare Ethics*, 1-11. https://doi.org/10.1017/S0963180124000045

Cambridge Quarterly of Healthcare Ethics (2024), 1–11
doi:10.1017/S0963180124000045

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Rights and Wrongs in Talk of Mind-Reading Technology

Stephen Rainey

Philosophy and Ethics of Technology Section, TU Delft, Delft, The Netherlands
Email: s.rainey@tudelft.nl

Abstract

This article examines the idea of mind-reading technology by focusing on an interesting case of applying a large language model (LLM) to brain data. On the face of it, experimental results appear to show that it is possible to reconstruct mental contents directly from brain data by processing via a chatGPT-like LLM. However, the author argues that this apparent conclusion is not warranted. Through examining how LLMs work, it is shown that they are importantly different from natural language. The former operates on the basis of nonrational data transformations based on a large textual corpus. The latter has a rational dimension, being based on reasons. Using this as a basis, it is argued that brain data does not directly reveal mental content, but can be processed to ground predictions indirectly about mental content. The author concludes that this is impressive but different in principle from technology-mediated mind reading. The applications of LLM-based brain data processing are nevertheless promising for speech rehabilitation or novel communication methods.

Keywords: brain data; chatGPT; fMRI; large language models; mind reading; reasons

Introduction

> …there is in principle no separating language from the rest of the world, at least as conceived by the speaker. Basic differences in language are bound up, as likely as not, with differences in the way in which the speakers articulate the world itself into things and properties, time and space, elements, forces, spirits, and so on. It is not clear even in principle that it makes sense to think of words and syntax as varying from language to language while the content stays fixed.[1]

Can signals recorded from the brain, like functional magnetic resonance imaging (fMRI) data, be fed into an large language model (LLM) like chatGPT to produce a mind-reading technology? Growing media attention would suggest that it could, if it cannot already.[2,3,4] But it would be hasty to jump to that conclusion. If true, and the mind was really legible in a direct way through technological means, there could be ramifications for all human beings. Ideas of the self, testimonial accuracy, privacy of thought, and other elements of being could all be challenged empirically. There could be grounds for

Earlier, upon its publication, I had the pleasure of reading the author's recent book, *Philosophical Perspectives on Brain Data*. In the book, the author critically examines the relations between brain data and claims about minds and brains in research and clinical contexts. The current article does an excellent job in following up on questions raised by the book. Specifically, he focuses on a timely case of applying a large language model (LLM) to brain data. Although experimental results appear to show that it is possible to reconstruct mental contents directly from brain data, by processing via a chatGPT-like LLM, the author argues that this apparent conclusion is not warranted and shows that they are importantly different from natural language. I enthusiastically endorse the publication of this article.

new rights or laws to protect novel vulnerabilities brought about by such developments. In medicine, psychiatry could be revolutionized where the mind could be read from captured brain signals. With such a development, a precision medicine approach to psychiatry would be possible, based on the convergence of neuroscience and artificial intelligence. This article aims to refute the idea that mental content is apt to be read in a way that words might be read off the pages of a book. This is important because ambiguity about purported mind reading technology will leave open the possibility for both over-reaction and complacency. Given the stakes in the area include ideas of the self, and including wide-ranging impacts in law and medicine, it is important not to do too much or too little in response. Rather, it is an essential first step to figure out exactly what the possibilities are in order to begin developing that response.

To focus this examination, the paper identifies claims concerning the application of an LLM to fMRI data, in a recent work from Jerry Tang *et al*.[5] Other studies seek to identify language from brain measurements using, for example, magnetoencephalography (MEG)[6] fMRI in combination with predictive models[7] or via electrocorticography.[8] In these and related studies, often the neural correlates of covert speech are sought. But in their article, Tang *et al* use an approach that can be taken as a form of mind reading technology—the apparent direct translation of data on brain activity to meaningful language with an LLM. It should be noted that the study's authors do not themselves make a direct claim of mind reading. In fact, in discussing "privacy implications," they stress that their approach requires participant cooperation as well as is being highly personalized to each participant. This would rule out a generalized kind of mind reading, of the sort anticipated in media reports and sci-fi stories. Moreover, as pointed out by Anne-Lise Giraud and Yaqing Su,[9] the use of fMRI in the methodology requires the use of a machine weighing some 5,000 kg so this is not a portable system. Nevertheless, Tang et al also write,

> This study demonstrates that the meaning of perceived and imagined stimuli can be decoded from BOLD fMRI recordings into continuous language[10]

and

> Results show that the decoded word sequences captured not only the meaning of the stimuli, but often even recovered exact words and phrases.[11]

This is in response to experiments in which experimental participants are listening to spoken words while an LLM is applied to real-time fMRI data. The "continuous language" produced in this way from the LLM appears to substantiate the correlation between brain data, as recorded by fMRI, and mental activity. This is contained in the claim that the meanings of perceived and imagined stimuli are recovered from the brain data. The claim is about mental content, and it is derived from neural data. I want to disagree with this claim and challenge the argument that is made for it. The claim, put more generally than above, is that technology-mediated mind reading is possible insofar as mental content can be gleaned from empirical observation of the brain. I argue that this specific convergence between neuroscientific investigation and an LLM does not indicate mind reading at all. Mental content is not recovered from empirical data.

The first line of challenge here is as follows: the assertion that the meanings of stimuli are recovered is erroneous. Meanings are not recovered from brain signal data but are only labeled. This does not detract from the impressiveness of the empirical correlation among heard words and GPT-generated outputs. The main part of the article will be pursuing this line of criticism. Responding to this, it might be objected that *of course* brain states correlate with mental states in significant ways. Any technology that can reliably identify brain states could, with appropriate sophistication and tuning, be used in some way to indicate mental states. Through examining associations among so indicated mental states, technology could play a mind reading role, even if it were not exactly of the sort typified in science fiction. In other words, without *de facto* mind reading, it could be possible nevertheless to yield some benefits of technology-mediated *pro tanto* mind reading. This is an obvious objection and one with clear pragmatic attraction. It will be addressed in the closing of the article.

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

Semantic reconstruction of continuous language from noninvasive brain recordings

The really significant aim of Tang et al is to try to recover directly the meaning of brain activity for a subject. As opposed to approaches investigating, for example, speech in terms of motor features, this decoding approach focuses on data from brain regions correlated with *semantic representations* during language perception, both real and imagined. A motor approach might allow for the reconstruction of speech by way of plotting articulatory motor dynamics such that a vocoder could be used to reproduce imagined speech.[12] In this kind of case, meaning is not derived directly from brain data but via the reconstruction of sounds themselves that are the bearer of spoken meanings. This articulatory motor approach is more like lip-reading than mind reading, in this sense. In their study, Tang et al explore the decoding of perceived and imagined speech stimuli based on blood oxygen–level dependent (BOLD) signal in brain regions associated with semantic representation. Despite apparent issues with the temporal resolution of BOLD signal and the rapidity of speech, Tang et al demonstrate decoding at the granularity of individual words and phrases. This relies on two main elements: a chatGPT-like autoregressive model to generate novel sequences of meaningful words and a "beam search algorithm" for efficient identification of probable sequence patterns among sequences or "continuations" as they are styled.

The approach uses a GPT-like interface that predicts likely word sequences, as a way to account for the low temporal resolution of BOLD data:

> …an impulse of neural activity causes BOLD to rise and fall over approximately 10 s. For naturally spoken English (over two words per second), this means that each brain image can be affected by over 20 words. Decoding continuous language thus requires solving an ill-posed inverse problem, as there are many more words to decode than brain images.[13]

To further optimize the approach, a "beam search algorithm" is deployed which allows the language model to produce predictions using previously decoded words as context:

> The encoding model then scores the likelihood that each continuation evoked the recorded brain responses, and the *k* most likely continuations are retained in the beam for the next timestep… This process continually approximates the most likely stimulus words across an arbitrary amount of time.[14]

In the training part of the work, seven participants underwent fMRI scans while listening to stories from *The Moth Radio Hour* and *Modern Love* podcasts. To associate meaning and visual perception, participants watched short clips from mostly wordless cartoons. Experimental tasks then included (i) imagining specific 1-minute portions of *Modern Love* podcasts, (ii) listening to different, simultaneous stories from *The Moth Radio Hour* while attending only to one, and (iii) paying attention to the events unfolding in the cartoons. From this, models of the participants' BOLD responses to many known verbal and visual stimuli could be constructed.

The language model, the GPT element for decoding brain data, was an advanced version "fine-tuned,"

> …on a corpus comprising Reddit comments (over 200 million total words) and 240 autobiographical stories from The Moth Radio Hour and Modern Love that were not used for decoder training or testing (over 400,000 total words).[15]

This model, using the additional element of the beam search algorithm would take the BOLD data produced during further experiment, listening to further as yet unheard podcast stories, and generate continuous text therefrom. The results of the testing phase are positively recorded by the researchers,

> Results show that the decoded word sequences captured not only the meaning of the stimuli but often even exact words and phrases, demonstrating that fine-grained semantic information can be recovered from the BOLD signal.[16]

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

This successful performance is also confirmed by a further validation step, wherein subjects exposed only to the decoded words could nevertheless relate much of the thrust of given stories, suggesting that meaning could be mediated through a perceiver, decoded via BOLD signal, and reproduced by a third party:

> We also tested whether the decoded words captured the original meaning of the story using a behavioral experiment, which showed that nine of 16 reading comprehension questions could be answered by subjects who had only read the decoded words.[17]

The cartoon paradigm is also interesting as it suggests perception and/or anticipation of responses to viewing films follows the plot to some appreciable degree, reflected in brain activity. This kind of response is implicated in considerations of theory of mind (ToM):

> When we watch movies, we consider the characters' mental states in order to understand and predict the narrative. …These results complement prior studies in adults that suggest that ToM brain regions play a role not just in inferring, but in actively predicting, other people's thoughts and feelings, and provide novel evidence that as children get older, their ToM brain regions increasingly make such predictions.[18]

This is no doubt a sophisticated and intriguing approach that aims to recover fine-grained data from slow brain signals. It looks on the face of it like a validation of at least some of the media claims about "mind reading," insofar as a machine appears to be taking semantic content directly from the physical activity of a subject's brain. But before going straight to that conclusion, it will be informative to take several detours through LLMs, and meaning and reference before utilizing some broader philosophical concepts in order to re-evaluate claims that this is actually what is happening.

Communication and interpretation

GPT applications like ChatGPT and the various versions of GPT1-4 are examples of LLMs. The "GPT" stands for *Generative Pre-trained Transformer*, which is a terminology evoking the nature of the model as having been trained on vast amounts of data, like human written text, such that it can *transform* its training data and *generate* what appear to be novel texts, among other things. This is worth pursuing a little further, in order to come to a view about what GPT can be said to be doing when it is producing its outputs. This will be important because, while Tang et al use fMRI data rather than text input, the modes in which GPT processes data will be central to understanding how its application to neurodata does not amount to mind reading.

LLMs in general work on the basis of an architecture including interaction between a language encoder and decoder, processing vast arrays of textual data. This processing includes deep learning algorithms that map out implicit structures in the data. For its part, the encoder in an LLM serves to codify general features of inputs relating to how different parts of the input relate. This could amount to mapping features of the input functionally equivalent to semantic relations. On a sentence-to-sentence level, this might be illustrated with "the cat sat on the…" and "mat" being associated, or "royalty sit on…" and "thrones" being associated. These are more statistically likely than, say, "the cat sat on the board of trustees," or "royalty sit on the shelf." More generally, two associations can be mapped such that features functionally equivalent to *allusion* or other more abstract phenomena can be seen to emerge. For instance, "royal" may also be encoded as having an allusive meaning to grandeur. I say "functionally equivalent" here as the nature of the encoder, being based on deep learning, produces its codifications based on statistical regularities rather than judgments about meaning. In this sense, the encoder produces a model of statistical dependencies among input features, which amounts to likely associations among words. It does not produce a map of relations among meanings but a model that reflects many of the features a map of meanings might include. The result of the encoder is something like vector semantics;

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

the length of the vector standing as a proxy for context and thereby allow for that sentence-to-sentence as well as more allusive generalizations. So "royal" might be closely associated with kings and queens, while more distantly with *regality* as grandeur.

The decoder produces outputs based on user prompts. The prompt may be any string of text that the LLM can continue in a variety of ways (e.g., continuing the prompt, answering a question asked in the prompt, analyzing a text provided with a request for summarizing).[19] Effectively, the decoder responds to whatever the prompt is by processing the products of the encoder and generating an output. It takes its cues from the vector-encoded weighted dependencies among inputs latent in the encoder's input representations, which are those relations among words and phrases. This decoding stage is somewhat like taking a journey with many signposts along the way at which further directions are given—the course is set at each signpost by the weight of advice given and the immediately prior signpost's position. So, at a signpost where "royalty" appears, yet there has hitherto been no reference to queens and kings, the otherwise less likely allusive meaning of "grandeur" might be the steer and so the decoder will take the allusive rather than literal path to the next waypoint. Given the length of vector from the encoder standing as something like context, here in the decoder, vector length could be said to stand for something like a proxy for attention—how much of which parts of a prompt the decoder "pays attention to" will serve to inform limits on the weights taken as workable for the required output. A prompt relating to cats and sitting will be more probably about *mats*, whereas a prompt about a spy and their sitting might more probably be allusive and relate to *information* or *sources.*

The interaction between encoder and decoder allows for massive simplification of what could easily become the intractable task of simulating natural language. The fact that language is ordered syntactically, grammatically, and conventionally means patterns can be readily identified by deep learning. Through encoding and decoding rather than representing the entire pattern, LLMs can streamline their input-output relations and make the impressive artifacts we are now familiar with. Central to this description of LLM function is the observation that statistical dependency among inputs is the fabric of encoding and the driving force of decoding. There is no rational dimension to this, except to the extent that "rationality" is captured in the language used as input data. The processing is data transformation, utilizing, for example, the mathematics of Euclidean space. While this is apt for data transformations, Mark Bishop highlights that it is fundamentally different from the kinds of relations that hold among objects in the real world:

> …the world, and relationships between objects in it, is fundamentally non-linear; relationships between real-world objects (or events) are typically far too messy and complex for representations in Euclidean spaces and smooth mappings between them—to be appropriate embeddings (e.g., entities and objects in the real-world are often fundamentally discrete or qualitatively vague in nature, in which case Euclidean space does not offer an appropriate embedding for their representation).[20]

Competent language users are normally capable of wading through the varied dimensions of semantic and pragmatics in their spoken or written language without overt appeal to the structures and patterns they use in order to produce meaningful speech action.[21] Speakers do not typically say what they say because there is a common precedent for saying what they say. Rather, they discern reasons and respond rationally (in the best-case scenario). This can include what on paper look like non-sequiturs. Perhaps you say "it's cold" and I interpret this as an instruction to close the window. As interpersonal behavior, this seems fairly straightforward to understand.[22,23] As a transformation in terms of data, it can only be modeled in terms of the frequency of one thing following the other. Even if the interpretation of your observation was not common, it would be easily understood why I closed the window given the observation. But an LLM might not associate the two things easily at all unless it became conventional and common for one thing to follow the other. Even in highly ritualized contexts, like legal proceedings or religious ceremonies where specific words must be spoken, the patterns' regular recital is not the speaker's reason for speaking as they do. Rather, their reasons come from wishing to participate in legal proceedings or religious ceremonies. This is a contrast between natural language and LLM outputs,

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

between statistical regularity and data transformation as explanation for LLM output, versus reasons-based activity for speakers.

Interpolation of Tang and philosophy of communication

Whereas Tang et al write "This study demonstrates that the meaning of perceived and imagined stimuli can be decoded from BOLD fMRI recordings into continuous language," it does not clearly do this. The meaning of the stimuli is not given but is labeled. That is, the *significance* of the content is not available from the recording, though the object of brain responses is labeled. In this instance, GPT is applied to fMRI data associating brain states with words, phrases, or expressions in a large text database. When it gets a likely "hit," it goes on to associate that hit with tropes and motifs from the database. It does this continuously, outputting intelligible text aiming to reproduce mental content faithfully from the brain activity of participants. But this is as much about intelligibility of output from corpus tropes as it is brains.

Neither mental nor semantic content is accessed, but the objective brain response to some stimulus is labeled in this approach. This is something different, not unimpressive by any means, but a different outcome. This can be further seen with the description of the perceived speech experiment, in which subjects listened to extracts from podcasts. The language model had been trained on other examples of the podcasts in question but not the specific episodes used in the experiment. But we cannot play down the effects of genre or editorial style, wherein outputs are finessed toward uniformity. Different shows have different styles that, while sometimes tricky to define, are nonetheless recognizable. These constraints can only serve to boost the predictive success of a language model through shortening the vector distances among model elements. What we can observe here is the pattern-matching function of an LLM applied to brain response and how it constrains forward brain state changes. This can be seen in light of LLM function in general and associated with the specifics of a language. But it is not directly operating with *meanings*. We are seeing the labeling of brain states and their dynamics in light of constrained language model dynamics.

When Tang et al suggest that, "Given any word sequence, this encoding model predicts how the subject's brain would respond," a few caveats are missing. The brain state dynamic model, in itself predictable given something like a movie experiment on the basis of ToM findings, can be associated with a highly constrained LLM approach in order to present correlations among the ToM findings and the dynamics of a language model. In other words, there is the possibility to label dynamic brain state changes in response to experimental stimuli using an LLM. This is not the same as stating that "Since our decoder represents language using semantic features rather than motor or acoustic features, the decoder predictions should capture the meaning of the stimuli."[24]

Something similar can be seen in the cartoon experiment too. There is not so much variety in the narrative structure and content of cartoons, especially those from Pixar. The predictive space here is also very narrow. We could say the system is responding to and labeling the ToM prediction dynamics of the cartoon viewer as the cartoon itself. Moreover, in having been trained specifically on the responses of the viewer with respect to large bodies of text, there will be material available for labeling the brain states recorded and their dynamics. The leap to talk of meaning is premature. It is not justified to conclude that meanings of stimuli can be decoded. This point about labeling is an important one worth pursuing a little more.

Pointing

It is often a successful strategy to point at something in order to clarify an ambiguity. For instance, if you ask me what "konijn" is when I talk to you in Dutch, I might wait until a rabbit appears in our surroundings and then point at it. I could say something like, "That's a konijn." You would probably, in those circumstances, then come away with a good idea of what konijn meant, that is, "rabbit." Part of why that works in general is that something zipping around against a static background is likely to be seen as a salient part of the environment. The salience of the thing zipping around will also include the implicit

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

idea that it is indeed one objective thing, otherwise its zipping around would not make much sense. So when, in our shared space, out pops a konijn or a rabbit, we do not have to do much work to recognize that it is that one thing to which we are referring in either case.

Philosophically speaking, this approach to language has limits.[25] If we were to follow through with it as a general account of meaning, we would hit puzzles and problems. The issues would cluster around the use of meaning as reference. This is just to say that conflating the thing talked about with the words used to talk about it is not a sustainable way to analyze meaning. Think of Clark Kent and Superman. They are the same reference, that is, the same person is both Clark Kent and Superman. If you point at Clark Kent, you point at Superman. But consider:

> Clark Kent disappears from the crowd of onlookers at a moment of emerging crisis. Superman swoops from the sky and saves the day.

This would be a standard bit of Superman-style action. Given both Clark Kent and Superman are the same person, and assuming meaning is reference, then we should be able to swap the names around and maintain the meaning of the text. But this clearly does not work. Clark does not swoop, and Superman was not in the crowd so could not disappear from it. The difference between them comes down to things an audience do or do not know about Clark Kent or about Superman. While it is true that they are one and the same reference, nonetheless "Clark Kent" or "Superman" might be good or bad ways to refer to some set of circumstances.

Not even for those who know about Clark/Superman's dual identity would pointing serve to work as picking out meaning. Pointing at either persona as a means of identifying could only serve as pointing did to correlate konijn and rabbit along with hefty background conceptualization. At least, pointing at one of the personas and uttering its name would be relative to time and circumstance and revisable in the event of changed context. In other words, we cannot collapse meaning to reference even in contrived circumstances: meaning requires attention to wider resources like context and change.

Rather than two people communicating and pointing at objects in the shared environment, now imagine a person whose interlocutor is an LLM and it is processing fMRI data recorded from the person's brain. What should we think when the person looks at Clark Kent, and the model translates from their brain "I see Clark Kent?" It would be impressive that a technology could read off such meaningful content from mere data. What if the model translated the brain data as 'I see Superman?" Would the output be false? (Table 1).

The sameness of Clark Kent and Superman is a matter of knowing under which aspects of identity one or the other name is relevant. The language model lacks the information from the brain data alone. Whether we considered its outputs as true or false, there is no way to ascribe rightness or wrongness to the model itself in any such case since it lacks the conceptualization of "knowing under an aspect" to be able to apply inferences with respect to it.

If the LLM correlated "Clark Kent" with "Superman" based on enough similarity in overall brain activity from the brains of those in the know about the dual identity, it still would not have the identity relation. This is because the language model process, however complex, is essentially a means of pointing to identify meaning. It cannot conceptualize. For instance, let us say I know about the dual identity. Back at the scene where Clark has disappeared from the crowd, but before Superman has swooped in, I could look at the gap in the crowd and know Superman is about to arrive. My brain activity would be that of someone looking at a crowd of onlookers. It might be odd, but it might also be feasible, to suggest that if I were talking with another person *in the know* about Clark/Superman, we could point at the gap and meaningfully say, "That is Superman." But this should serve to indicate the large amount of context required in ascribing meanings to objects of experience, including gaps. Pointing alone will not do.

This idea of pointing is akin to what happens when an LLM labels a portion of brain activity data as meaning one thing or another. Current brain data are the prompt that, based on modeled features from the corpus and some immediate priors, triggers the LLM to produce a specific output and onward prediction. Just as an LLM might produce predictions about Superman and Clark Kent such that they could be interpreted as being the same person, but without understanding anything about the identity

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

Table 1. A table suggesting some possible queries relating to how GPT outputs might be interpreted by an audience, when an issue of context, change, or wider knowledge is at stake

GPT output	True, false, something else	Detail
I see Clark Kent	True	Clark Kent is dressed as Clarke Kent, doing Clark Kent things like being a journalist
	False or something else?	The person knows about the dual identity, so they see Clark Kent, but know too that they see Superman
I see Superman	False or something else?	Superman is not present, Clark Kent is. He is dressed as Clarke Kent, doing Clark Kent things like being a journalist
	True	The person knows about the dual identity, so they see Clark Kent, but know too that they see Superman

Note: The way the brain responds to stimuli associated with Clark Kent is connected statistically by the GPT-like application with training data about Clark Kent. That is, the brain's activity when confronted with the visual image of a bespectacled reporter and evaluative judgments about bumbling behavior correlate with mentions of Clark Kent in texts, and spoken references in podcasts. The way the brain responds to stimuli associated with Superman,meanwhile, is connected statistically by the LLM with training data about Superman. That is, the brain's activity when confronted with the visual image of a caped figure and evaluative judgments about strength are associated with mentions of Superman in texts, and spoken references in podcasts.

relation, so too labeling brain activity as part of continuous language output. Nowhere in this labeling does meaning arise. This can be applied to the experimental work by Tang et al.

In the experiments referred to above, while an experimental participant is listening to words being spoken aloud, saying:

> …I had no shoes on I was crying I had no wallet but I was ok because I had my cigarettes…[26]

The LLM applied to their fMRI data in this instance outputs the following text:

> …we got in my car and I was crying I didn't have my purse I don't have any money to pay for gas…[27]

This is impressive—it picks out "crying" as featuring in the story and that the narrator is describing some sort of lack (shoes, not purse and money). But it is not the story. This might tell us that, aside from the brain states evoked from the perceived story, in the large training corpus, "crying" is associated with lack of some kind. This would be a statistical correlation in text, not LLM-detected mental content. At each moment of the language model's application to fMRI data, it is effectively pointing at a word sufficiently correlated with the brain state—based on prior training specific to a participant—such that it might be the one currently perceived. We should think of the role of the LLM in this way: as pointing at words in its database correlating with brain states evoked in response to known stimuli and predicting likely next words according to a database of texts. More than this, though, it applies in each case a statistical process to predict the next words given the last. It does not just point out but points ahead too. This is part of what makes it so impressive. But it still does not get beyond meaning as reference. It makes intelligible outputs because it has a large corpus of intelligible text, but the recovery of meaning does not arise as this is bound to the corpus, not the participant's mind. This use of a predictive algorithm based on text is already enough to rule out "mind reading" according to Charles Rathkopf et al whose careful analysis specifies:

> Neuroscientific mind reading is (i) discerning mental content (ii) from a prediction about some property of an experimental condition, where the prediction (iii) does not, during testing, capitalize on the intentional production of conventional symbols by the subject and (iv) is computed by a prediction algorithm that takes exclusively neural data as input.[28]

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

Owing to the use of a text corpus in the LLM approach, a mind/meaning complication arises concerning semantic indeterminacy or epistemic sufficiency. Consider, in terms of the above two sentences, what we could think if the LLM had output:

- I lost all my things, I was upset, but I felt all right.

We might say this is close enough, that it was a good output, or that it matched the stimulus meaning fairly well. But this judgment of adequacy reflects the malleability of language, not the performance of a system extracting meaning from brain data. The behavioral dimension of the experiments by Tang et al seems relevant to this point. That third parties were able to generally understand stories as related via LLM-produced text leans heavily on this phenomenon of language, rather than on the accuracy of mental content recovery. The phenomenon is intralinguistic.

A critical point might be made at this juncture as follows: "Isn't prediction enough to warrant claims for mind reading? After all, if accurate guesses can be made in a reliable way isn't this *basically, pretty much* mind reading?" This could be called a *pro tanto* objection, in that the outputs from the LLM in this kind of experiment could be seen as mind reading *to some extent*, or to some meaningful extent. My thought here is that it is significantly different, even if there might be examples where prediction and "reading" do not appear to be substantially different. I want to suggest that this relates to how we describe the dynamics of mental changes versus those of brain states.

Mental content can change with respect to a variety of causes, including perceptions, mood, spontaneous and unidentified factors. But an important mode of change it can undergo relates to rational change. We can, or we ought to, change our minds according to *reasons.* Brain states, on the other hand, change according to physical processes capturable by recordings like BOLD signal, MEG, EEG, and so on. Mental content gains significance by way of its rational connections, and brain states are interesting because of their neurochemical and other properties. A series of brain states, labeled and associated with language dynamics implicit in an LLM, remain a set of physical descriptions connected by way of statistical models. Even when these correlate closely with a set of avowed mental contents, the mental contents gain their significance by way of the rational connections and not the statistical. In the last snippets just mentioned, the speaker says they were ok *because* they had their cigarettes. This is not because having cigarettes is statistically associated with being okay but because this person's narrative included that detail of their experience. In the LLM processed output, however, this is all that can be said —crying, lacking a purse and money are presented together because they co-occur in the right way, given the training corpus, to be output.

Brain states can be predicted, based on prior brain states and activation data yielded from stimulus conditions in the training phase of experiment. These activation data are labeled by the machine learning process, and decision boundaries are set such that they effectively act as prompts in the LLM procedure outlined above. But though the labels applied to each state are themselves meaningful in being derived from the linguistic corpus, the predictions do not amount to reading mental content. Nevertheless, the intuitive appeal of this "near as makes no difference" sort of observation ought not to be overlooked. Deliberate misuse or mistaken faith in the reliability and accuracy of this kind of approach, fusing neurodata with LLMs, could lead to bad situations across contexts like law or psychiatry wherein "hidden thoughts" might be of particular interest. While the technology itself is not there for mind reading, the wider systems in which it appears need to be alive to the idea that it may be taken for real mind reading technology.

Conclusions

The use of LLM technology applied to fMRI data was claimed to recover meanings of stimulus from experimental participants listening to podcasts or watching cartoons. Such claims appear to ground speculations in the media about mind-reading technology. But looking more closely at how LLMs work, and at some features of language and meaning, the experimental results appear to show instead a way of

https://doi.org/10.1017/S0963180124000045 Published online by Cambridge University Press

labeling brain states in response to stimulus. It seems to be a highly complex form of technological pointing and labeling. This can ground some predictions about mental states. Nevertheless, this is not mind reading.

The quote from philosopher Willard Van Orman Quine at the opening of the article indicates the complexity of language when considered as a repository of meaning as reference. Synonymy might be taken as a phenomenon of simply swapping one word for another, like using a thesaurus. But Quine's analysis suggests that word-for-word translation or transliterations miss the speaker perspective. Words thought of as labels on objects apt to be pointed out, or swapped around, do not account for speakers' meanings, or the holistic rational relations that hold between one word and all the others. Mental states, enjoying rational relations that have significance for the speaker or the experimental participant, are only alluded to by LLMs whose 'language' is a statistical transformer model based on a massive text corpus. Brain data, recorded by fMRI and processed by an LLM application, can best be considered synonymous with mental content in an allusive or a metaphorical sense. An instance of fMRI data might be acceptably close to an avowed (albeit self-reported) mental state such that predictions can be made about onward states. But this is not mind reading as much as it is an indication that language is often closely related with thought. Moreover, in the case discussed here, fMRI data are essentially a prompt for an LLM. The result of this being that the eventual output is as produced as rationally as any LLM-produced text, which is to say *not at all.*[29]

As a noninvasive way to produce text from brain signals, the kind of technology here seems promising, perhaps most importantly for speech rehabilitation or novel communication application contexts. In psychiatric contexts, this technology might have some uses, but by no means could be considered a reliable diagnostic tool. Perhaps, as with some instances of chatGPT use, outputs in this context could be used in psychiatric contexts as prompts for discussion between therapist and patient. But the attendant risks in the possibility for producing inappropriate outputs would likely outweigh any benefits. As for the need for new legal protections, this kind of application likely ought to prompt discussion. While for now the exact iteration of the technology is here unwieldy and not portable at all, the principle could be finessed into more agile systems. In the analysis, the *pro tanto* objection was dismissed on the basis that prediction and mind reading are not the same thing. But this significant difference does not rule out intentional misuse or use in ignorance of shortcomings in applications claiming to reveal a subject's innermost thoughts. Clear messaging ought to be attached to work on this kind of technology emphasizing its fallible, predictive nature. The popular media, and other commentators, would thus do well to reassess mind reading claims.

Notes

[1] Quine WVO. *Word and Object*. Cambridge, MA: MIT Press; 1960, at 61.

[2] Devlin H. AI makes non-invasive mind-reading possible by turning thoughts into text. *The Guardian* 2023 May 1, Section Technology; available at https://www.theguardian.com/technology/2023/may/01/ai-makes-non-invasive-mind-reading-possible-by-turning-thoughts-into-text (last accessed 21 November 2023).

[3] Whang O. A.I. is getting better at mind-reading. *The New York Times* 2023 May 1, Section Science; available at https://www.nytimes.com/2023/05/01/science/ai-speech-language.html (last accessed 21 November 2023).

[4] Wilkins A. Mind-reading AI works out what you are thinking from brain scans. *New Scientist* 2022 Oct 14; available at https://www.newscientist.com/article/2342509-mind-reading-ai-works-out-what-you-are-thinking-from-brain-scans/ (last accessed 21 November 2023).

[5] Tang J, LeBel A, Jain S, Alexander H. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* 2023;**26**(5):8583. doi:10.1038/s41593-023-01304-9.

[6] Dlexandr A, Caucheteux C, Rapin J, King J-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* 2023;**5**(10):1097–107. doi:10.1038/s42256-023-00714-5.

7. Grandchamp R, Rapin L, Perrone-Bertolotti M, Pichat C, Haldin C, Cousin E, et al. The ConDialInt model: Condensation, dialogality, and intentionality dimensions of inner speech within a hierarchical predictive control framework. *Frontiers in Psychology* 2019;**10**:2019. doi:10.3389/fpsyg.2019.02019.
8. Pei X, Barbour DL, Leuthardt EC, Schalk G. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of Neural Engineering* 2011;**8** (4):046028. doi:10.1088/1741-2560/8/4/046028.
9. Giraud A-L, Su Y. Reconstructing language from brain signals and deconstructing adversarial thought-reading. *Cell Reports Medicine* 2023;**4**(7):101115. doi:10.1016/j.xcrm.2023.101115.
10. See note 5, Tang et al. 2023.
11. See note 5, Tang et al. 2023.
12. Bocquelet F, Hueber T, Girin L, Savariaux C, Yvert B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLOS Computational Biology* 2016;**12**(11): e1005119. doi:10.1371/journal.pcbi.1005119.
13. See note 5, Tang et al. 2023.
14. See note 5, Tang et al. 2023.
15. See note 5, Tang et al. 2023.
16. See note 5, Tang et al. 2023.
17. See note 5, Tang et al. 2023.
18. Richardson H, Saxe R. Development of predictive responses in theory of mind brain regions. *Developmental Science* 2020;**23**(1):e12863. doi:10.1111/desc.12863.
19. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 2020;**30**:681–94. doi:10.1007/s11023-020-09548-1.
20. Bishop JM. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology* 2021;**11**:4; available at https://www.frontiersin.org/articles/10.3389/fpsyg.2020.513474 (last accessed 20 March 2023).
21. Grice HP. Logic and conversation. In: *Studies in the Way of Words*. Cambridge, MA: Harvard University Press; 1995:22 pp.
22. Grice HP. Meaning. *The Philosophical Review* 1957;**66**(3):377957. doi:10.2307/2182440.
23. Hare RM. *The Language of Morals*. Clarendon Paperbacks. Oxford: Clarendon Press; 2003, *passim.*
24. See note 5, Tang et al. 2023.
25. See note 1, Quine 1960, at chap. 2.
26. See note 5, Tang et al. 2023.
27. See note 5, Tang et al. 2023.
28. Rathkopf C, Heinrichs JH, Heinrichs B. Can we read minds by imaging brains? *Philosophical Psychology* 2023;**36**(2):221023, at 233. doi:10.1080/09515089.2022.2041590.
29. See note 18, Richardson, Saxe 2020.

Cite this article: Rainey S (2024). Rights and Wrongs in Talk of Mind-Reading Technology. *Cambridge Quarterly of Healthcare Ethics*: 1–11, doi:10.1017/S0963180124000045