Can we read minds by imaging brains?*

Charles Rathkopf
Jan-Hendrik Heinrichs
Bert Heinrichs
Institute for Neuroscience and Medicine
Jülich Researche Center†

Abstract

Will brain imaging technology soon enable neuroscientists to read minds? We cannot answer this question without some understanding of the state of the art in neuroimaging. But neither can we answer this question without some understanding of the concept invoked by the term "mind reading." This article is an attempt to develop such understanding. Our analysis proceeds in two stages. In the first stage, we provide a categorical explication of mind reading. The categorical explication articulates empirical conditions that must be satisfied if mind reading is to be achieved. In the second stage, we develop a metric for judging the proficiency of mind reading experiments. The conception of mind reading that emerges helps to reconcile folk psychological judgments about what mind reading must involve with the constraints imposed by empirical strategies for achieving it.

*This is a pre-print of an article that appears in *Philosophial Psychology* in February of 2022.

†We would like to thank audiences at the Dutch Distinguished Lecture Series in Philosophy and Neuroscience, the Italian Association of Cognitive Science, and the University of Mississippi. Thanks also to two anonymous referees for extensive and helpful commentary, and especially to Daniel Dennett and Rosa Cao for their insightful feedback on earlier drafts.

Contents

1 Introduction

In recent years, the term "mind reading" has come to be used to describe a family of brain imaging experiments aimed at decoding patterns in neural data. The term can be found, not only in publications aimed at a popular audience (Intagliata, 2008; Poldrack, 2018; Wilson, 2019), but also in mainstream neuroscience journals (Norman et al., 2006; Hasegawa et al., 2009; Reddy et al., 2010; Roelfsema et al., 2018). Typically, the term "mind reading" is used to denote a certain achievement that has not yet been made, but which is likely to be made in the near future, given the pace of innovation in imaging technology (Peckham, 2011). This usage raises an obvious pair of questions: (i) what exactly would the predicted achievement look like? And (ii) how will we recognize it when it comes? Although there is a large and growing literature on how decoding experiments should and should not be interpreted, there are, as far as we can tell, virtually no published attempts to provide rigorous answers to our two questions.[1] Perhaps this is because those in position to provide the answers are inclined to view claims about the imminence of mind reading technology as mere hyperbole. Virginia Hughes, a reputable science journalist who appears disappointed with her peers for what she takes to be their gullibility in the matter, writes that claims of having achieved neuroscientific mind reading are laughably false" (Hughes, 2011). Similar sentiments are easy to find elsewhere.[2]

Unfortunately, neither the boosters nor the skeptics have made systematic attempts to provide evidence for their claims. In fact, given the rather loose collection of associations evoked by the term "mind reading," it is not clear what kind of evidence would constitute support for either position. Until we have some understanding of what the predicted achievement would involve, questions of evidential relevance are destined to go unanswered. Boosters and skeptics should agree, therefore, that there is pressing need for a rigorous analysis of what it takes to read a person's mind. That is the task we propose to undertake here.

In what follows, we first develop a qualitative analysis of mind reading that articulates a set of individually necessary and jointly sufficient conditions for reading a mind. We then extend that analysis by developing a metric for comparisons between individual episodes of mind reading with respect to their de-

[1] One exception to this is a brief but fascinating discussion in the opening paragraphs of Tong and Pratte (2012). This passage is discussed below.

[2] See, for example, Chou (2012).

gree of proficiency. Together, these analyses permit us to reflect systematically on how far mind reading technology has come.

2 Just a matter of definition?

We begin our investigation with the most rudimentary possible strategy - consulting a dictionary. According to Dictionary.com, mind reading is "the ability to discern the thoughts of others without the normal means of communication, especially by means of a preternatural power." If we set aside the distracting qualification about preternatural power, this definition is not a bad start. It takes two to communicate, and so, as we will see in more detail below, ruling out the normal means of communication involves placing restrictions both on the activity of the subject whose mind is to be read, and on the activity of the scientist who hopes to do the "reading." Nevertheless, despite having prompted this insight, the dictionary definition is rather thin. It does not put us in position to judge whether any contemporary imaging experiments have succeeded at reading minds, or whether more sophisticated variants of those experiments are likely to succeed in the future.

How then, should we proceed? And what kind of expertise is required for the job? To answer these questions, it is important to note that the term "mind reading" - in the sense in which we intend to use it - is not a theoretical term. Its meaning is not defined by the inferential role it plays in some established body of scientific theory. It is a folk-psychological term for a phenomenon that, until recently, was far more likely to appear in a science fiction novel than an academic journal. Moreover, where the term *has* been used in scientific journals, its meaning has not been regimented with technical definitions. Instead, the meaning of the term has its roots in untutored intuitions about what minds are, and about how thoughts are stored and transmitted.

This suggests that the task of articulating the meaning of the term "mind reading" falls at least partially outside the purview of cognitive neuroscience. But only partially. Folk-psychological concepts are sometimes corrupt, and fail to refer to anything real. We hope to describe a phenomenon whose existence is at least logically compatible with our best, empirically-grounded science of the human mind. It will be necessary, therefore, to consult that science as we construct our explication.

Because our project seeks to integrate folk-psychological thinking with with empirical science, it is more synthetic than analytic. The goal of a synthetic philosophical project is not to construct a compact definition, of the sort one might see in a dictionary. The goal is rather to construct an *explication* (Carnap, 1962). An explication articulates what a concept *ought* to mean, given the concerns associated with the phenomenon to which it purports to refer. It is an attempt to improve a concept, rather than to convey an established concept meaning.

What is required of a good explication? One central requirement is that it must retain some grounding in our pretheoretical concern with the focal topic. In the absence of such grounding, explication devolves into mere stipulation. We must ask ourselves, therefore, why the prospect of neuroscientific mind reading *matters*. Or, to put the question slightly differently: what drives the widespread fascination with the possibility of reading minds with brain imaging devices? Some academic discussions of the topic are ostensibly motivated by an ethical concern that mind reading technology will threaten the privacy of thought (Roelfsema, 2019; Roskies, 2014). Threats to privacy may be part of what fascinates (or worries) many people, but we doubt that it is the central factor. After all, participation in mind reading experiments is voluntary. Moreover, given the cumbersome nature of brain imaging equipment, there is little reason to fear attempts at covert data collection.

In our view, mind reading technology is primarily fascinating because it promises to circumvent what psychologists and philosophers have long viewed as a fundamental constraint on the acquisition of knowledge about other minds. Traditionally, acquiring *detailed* knowledge about the minds of other people has necessitated that those people speak, write, or sign. In each case, information is encoded in a public symbol system in which symbol meanings are underwritten by social convention. Mind reading is primarily fascinating, we submit, because it *appears* to be a method of interpreting people that does not rely on the use of conventional symbols, and which, therefore, appears to provide more direct epistemic access to the contents of another person's thought.

2.1 Conventional symbols

Given this hypothesis about the source of fascination with mind reading, we must include in our explication a necessary condition of the following sort: an experiment will count as mind reading only if it manages to discern mental

content without relying on the use of conventional symbols. As it stands, this condition is too broad. We do not want to require, for example, that the neural activity supporting an episode of mind reading be completely free of influence from conventional symbols. Both human evolution and human development occur in social contexts shot through with conventional symbols, and so we are unlikely to be able to identify any determinate bout of neural activity that has not been, in some sense or other, influenced by them (Seligman et al., 2016; Rathkopf, 2021).

The phenomenon we want to rule out is not, therefore, the influence of conventional symbols per se. It is rather a particular kind of reliance on conventional symbols; one in which they are purposefully used to transmit information about mental content. The most obviously unacceptable case is one in which the subject simply uses natural language to describe what they are thinking. We would like a criterion that rules out this case, but also rules out more subtle variants of the same communicative phenomenon. When we talk, we not only *use* conventional symbols; we *intentionally produce* concrete tokens of conventional symbols. It is this intentional production of conventional symbols that must be avoided.

There is another sense in which this condition threatens to be too strong. To discover associations between neural activity and mental content, it may be necessary to ask an experimental subject to speak, in order to confirm that the experiment is running as intended, for example, or that the task has been understood properly. Here we must briefly anticipate the discussion in Section 3 by noting that mind reading experiments typically rely on machine learning, and can, therefore, naturally be divided into training and testing phases. The training phase typically relies on a form of supervised learning, in which neural data is labeled. Since these labels are themselves conventional symbols, the activity in the training phase will not count as a mind-reading. If mind reading occurs at all, it occurs in the testing phase, after the associations between neural activity and mental content have been learned. This narrowing in the scope of the restriction, moreover, allows us to be correspondingly permissive about the use of behavioral data in the training phase, and in particular, allows us to accept the use of verbal report.

Here then, is the official version of our first necessary condition: in order for an inferential process to count as mind reading, it must not rely on the intentional production of conventional symbols by the subject during the testing phase of the experiment.

This condition rules out at least two interesting cases that we might otherwise be obliged to include as forms of mind reading. The first case concerns non-invasive silent speech interface devices, such as the *Alterego* device recently developed by the MIT Media Lab (Kapur et al., 2018). The purpose of such devices is to increase the speed of text-based communication by bypassing the relatively slow thumbwork required to type out symbols on a smartphone. These devices are marketed as a form of "thought-to-text technology" - a description that sounds quite close to mind reading. However, these devices produce written text by interpreting myoelectric signals in the face and neck that are produced by subvocalization. Since subvocal speech is conveyed by means of conventional symbols, our first condition rules it out. Note that such devices are not ruled out merely because myoelectric signals manifest themselves on the wrong side of the brain-body barrier. One can imagine a similar technology designed to interpret signals in the supplementary motor area (SMA) that encode the muscle commands that will subsequently be translated into speech. As long as these decidedly *neural* signals are subject to intentional control, this variant of the *Alterego* device would be ruled out as well.[3]

The second case is perhaps more surprising, since it involves genuine brain imaging. Patients suffering from the advanced stages of amyotrophic lateral sclerosis (ALS) lose virtually all capacity for motor control, and, as a result, lack the means by which to communicate. One potential solution to this problem is to use brain activity itself as a vehicle of communication. By using functional near infrared spectroscopy (fNIR) targeting Broca's area, which is differentially active whenever effortful linguistic tasks are performed, it may be possible for patients to use the magnitude of their own neural activity as a kind of binary coding channel. For example, if a patient counts backwards from 100, the level of activation in Broca's area will increase. If the patient instead imagines a placid lake, that above-baseline activity will dissipate. Doctors and patients can agree to let the increased activation level code for "yes" and the baseline activation level code for "no." In the absence of other medical complaints, the patient could then answer any yes/no questions a doctor might wish to pose. Progress on fNIR communication for ALS patients has been slower and less successful than initially anticipated.[4] Nevertheless, for our purposes, the practical viability

[3]How to tell which neural signals are intentional is of course an open question, about which we would like to remain neutral.

[4]The reputation of this work has also been tarnished by recent allegations of scientific misconduct (Vogel, 2019).

of the method is less important than the possibility illustrated by its design. For our purposes, the crucial point about this case is that, by establishing a binary code, the patient/doctor team bestows a conventional meaning upon patterns of brain activity, and the patient intentionally exploits that conventional meaning to transmit information about their mental state.

2.2 Neural data and quotidian interpretation

The dictionary definition with which we began suggested that we must rule out the normal means of communication. As noted above, communication has both a productive side and a receptive side. Our anti-convention condition is designed to restrict activity on the productive side. The goal of this section is to formulate an analogous condition for the receptive side. On a first pass, we might say that mind reading cannot rely on the quotidian interpretive skills of the researcher. We will now try to make this condition more precise.

We begin with a terminological observation. "Mindreading" is often used in cognitive and developmental psychology to denote the species-typical, non-technological capacity to interpret the mental states of others. Let us reserve the two-word term "mind reading" to distinguish the phenomenon at issue here from mindreading in the theory of mind sense. We can then ask about the relationship between these two concepts. The most salient difference concerns the input type appropriate to each. Our species-typical interpretive capacity takes speech, gesture, and other outwardly perceptible forms of behavior as input. Neuroscientific mind reading, by contrast, takes neural activity itself as input. If a purported episode of neuroscientific mind reading relied too closely on the quotidian mindreading capacities of the researcher(s), we would have grounds to discount it. This thought yields a slightly more refined version of the condition: in order for an inferential process to count as an instance of neuroscientific mind reading, it must take exclusively neural data as input.

This version of the condition remains imprecise, because we have not yet said anything about the referent of the term "inferential process." If we conceptualize the inferential process broadly, so as to include the design of the experiment, the collection of data, and all subsequent data analysis, then the restriction to neural data as input will preclude many commonplace and probably essential aspects of brain imaging research, such as examining the behavioral data to see whether the experimental task had been performed properly. We can narrow the condition further by distinguishing between the scientific experiment

broadly conceived, and the algorithmic component of the experiment in which a prediction is actually computed. It is only input to the predictive algorithm, which is employed during the testing phase of the experiment, that should be subject to the restriction. This narrowing of the condition yields two benefits. First, the researchers are free to reason freely on about all subject behavior, including verbal behavior. Secondly, the learning algorithm (e.g. backpropagation through a neural network), which is used during the training phase of the experiment, can be fine-tuned, if necessary, on the basis of behavioral data. Here then, is the official version of our second condition: in order for an experiment to count as an instance of neuroscientific mind reading, the input to the predictive algorithm must be restricted to neural data.

This condition is not trivial. Imagine an object recognition experiment in which the familiarity of the object type is correlated with the latency of the response (how long it takes the subject to press a button to record their judgment about a particular trial). In such a case, the response latency provides predictive information about the type of stimulus presented on each trial. Predictions that rely on such behavioral information, even if only in part, will be ruled out by our explication.

We have now described two necessary conditions on what it takes to read a mind. In practice, the two conditions overlap substantially, because they both rule out verbal communication between researcher and participant during the testing phase, and because verbal communication is the most salient kind of communication that would clearly undermine claims of mind reading. Nevertheless, the two conditions are worth distinguishing. Ruling out the production of conventional symbols delimits the contribution of the participant. Ruling out quotidian interpretive skills delimits the contribution of the researcher(s). Moreover, the two conditions are logically separable. A facial expression of fear provides non-neural data about one's mental state that is, arguably, free from convention. And our ALS case shows that it is possible to discern mental content by means of a predictive process that is driven by neural data, but which, nevertheless, does rely on convention.

There is another sense in which our two necessary conditions are intimately related: both are motivated by folk-psychological considerations. To refine and eventually complete our explication of mind reading, we must go beyond the insights folk psychology has to offer. And to do that, we must have at least a rough account of relevant experimental ideas. Developing such an account is the task of the following section.

3 Some elements of a generic mind reading experiment

Typically, mind reading experiments employ a family of techniques called multivariate pattern analysis, or MVPA (Haxby, 2014). These techniques begin with neural activation data, corresponding to a large number of measurements of neural activity. In fMRI experiments, each measurement corresponds to a particular voxel. In singe-cell electrophysiology, each measurement corresponds instead to one particular neuron. Regardless of the particular measurement technology used, each measurement site in the brain is represented as one dimension in a high-dimensional space, and each individual measurement represents a value for the dimension corresponding to the site from which it was taken. A pattern of activity at a time is then represented as a single point (or, equivalently, as a vector) in that high-dimensional space. Each point is associated with some property of the experimental condition in which it was elicited.

The experimental property in question might be stimulus category. Stimulus categories are associated with standard labels, such as "face" or "house." Once the data has been collected and labeled, a machine learning algorithm is employed to discover associations between the points, which represent neural activity patterns, and the labels, which represent the stimulus type that defines an experimental condition. For a given label, a learning algorithm draws a decision boundary around the points to which the label corresponds. Once that decision boundary has been constructed, it can be tested on new data, and measures of accuracy can be computed.[5]

Drawing a decision boundary like this is an example of stimulus classification. There are also more sophisticated methods, such as stimulus reconstruction. In recent years, it has become possible to reconstruct even complex stimuli, including video. The relationship between the neural data and the decoding prediction is often mediated by a deep neural network. The network is trained to extract feature vectors from training data, and the decoding model then searches for correlations between the neural data and those feature vectors (Wen et al., 2017; Shen et al., 2019).

Classification and reconstruction are both kinds of prediction that one makes within the context of a *decoding* experiment. Decoding experiments begin with

[5]This description is complicated by the use of cross-fold validation methods, in which a single data set is iteratively reused by splitting it into training and testing subsets, and choosing new subsets on each iteration (or "fold").

neural activation data, and then generate predictions about properties of the experimental condition in which that data was elicited. An *encoding* experiment does the opposite. It begins with knowledge of an experimental condition, combined with general knowledge about typical neural responses to conditions of that type, and then generates predictions about the pattern of neural activation that will be elicited in a new subject when they are confronted with the experimental condition in question. The capacity to make accurate predictions about neural activation data is undoubtedly impressive. However, since our goal is to discern mental content *from* neural data, decoding experiments are more directly relevant.[6]

4 A categorical explication

Many decoding experiments already satisfy the two necessary conditions developed in Section 2. The predictive phase of such experiments are not directly supported by conventional, symbolic communication. Moreover, they are driven by neural data in the relevant sense. What other conditions must be met, before we can say that genuine mind reading has been achieved? To help answer that question, we introduce a maximally simple experiment in which our two necessary conditions are satisfied, and use it as a test case.

Imagine a perceptual experiment in which participants are presented with a randomized series of pictures, half of which depict faces, and half of which depict houses. They perform a speeded one-back repetition detection task, in which a button is pressed to indicate whether the current image depicts the same house or same face as the one immediately preceding it. The participants achieve a high, but imperfect rate of success, which indicates that they performed the task attentively. Patterns of neural activity are recorded, and a model is trained on the data. In subsequent testing sessions, using the same participants and highly similar stimuli, the model correctly predicts stimulus type on over 90 percent of trials.[7]

The sort of predictive success achieved in such a spartan decoding experiment is underwhelming. Is there some additional ingredient we can add to the

[6] Nevertheless, encoding models are relevant to mind reading in at least two ways. First, a capacity for encoding is strong evidence that successful decoding is possible. Second, by building and testing encoding models, we acquire valuable information about how best to tune and improve decoding models. Hence, encoding and decoding models are mutually reinforcing.

[7] The experiment described here corresponds roughly to Haxby et al.'s landmark experiment published in Science in 2001.

recipe that will give us genuine mind reading? In one of the very few attempts by neuroscientists to provide a conceptual analysis of mind reading, Frank Tong and Michael Pratte (2012) refer to experiments of the sort just described as cases of mere *brain reading*. Mind reading, on their view, must satisfy two additional conditions. One of these conditions says that mind reading can only be achieved if the information recovered from the neural data is "fundamentally private and subjective" (Tong and Pratte, 2012, p. 485). But there is good reason to reject this condition. If a body of information is fundamentally private and subjective, there can be no empirical test that would confirm that the information had been decoded correctly. We can, of course, just ask the subject for a verbal self-report. However, in addition to precluding skeptical questions about the reliability of self-knowledge, self-report requires the subject to intentionally produce conventional symbols, and therefore violates our anti-convention condition.

The second necessary condition Tong and Pratte propose says that the experimental prediction must not have been achievable by means of simply glancing at the stimulus presented to the participant during the relevant interval. This condition is intuitively appealing, but also underdescribed. One interpretation of the condition is that, in order to count as mind reading, it must be literally impossible to guess the correct (label for) the triggered mental content, on the basis of having merely glanced at the stimulus. Interpreted this way, however, the condition is too strong. No matter how unusual a snippet of mental content might seem, lucky guesses remain theoretically possible. So this interpretation, much like the first condition Tong and Pratte proposed, renders mind reading impossible by definition. The natural response to the problem of lucky guesses is to soften the condition by making it probabilistic. One might formulate it as follows: in order for an experimental prediction to count as a mind reading, it must be the case that, in the absence of special knowledge about the experimental participant, the probability of an observer correctly guessing the participant's mental content, given only a glance at the triggering stimulus, is very low. The trouble with this interpretation is that it presumes the existence of a discrete probabilistic threshold that marks the divide between mind reading and brain reading, and it is hard to see what considerations could justify the selection of any particular threshold. If there is no non-arbitrary way to determine the hypothetical threshold value, then the probabilistic condition cannot

underwrite a qualitative distinction between mind reading and brain reading. [8]

Another way of interpreting Tong and Pratte's second condition is by focusing not on the difficulty of the guess, but on the manner in which the mental content is represented for the purposes of guessing. In the spartan decoding experiment described above, the prediction involves choosing the correct label for the stimulus-type from a list of labels. These labels play two roles: they index the stimuli for the purposes of experimental design, *and* they serve as the natural language representation of the targeted mental content. In light of this, one might be tempted to complain that the labels were not produced by the neural activity of the participant, and were instead *assigned* to the stimuli, from the outside, by the researcher(s). This assignment, the complaint continues, is based on a quotidian judgment about what people would typically think about a particular stimulus in an experimental context. Perhaps this is the heart of the problem. Perhaps what Tong and Pratte really want to insist on is that the label used to represent the targeted mental content must be generated by the participant, rather than assigned by the researchers.

This suggestion is more problematic than it first appears. Notice that we cannot rely on the participant to intentionally produce the phonetic components of the label. Doing so would, once again, constitute a violation of our first anti-convention condition. We must therefore hope to identify language-like labels in neural activity, which, despite being language-like, have yet to be intentionally produced in natural language format. The task would seem most practicable if it turned out that human cognitive architecture conforms to the language of thought hypothesis, according to which content is intrinsically sorted into discrete and determinate propositions in a mental language (Fodor, 1975). If the language of thought hypothesis were true, we might reasonably hope that mentalese words or sentences could, in some sense, be isolated and decoded *directly*. However, notice that even if we manage to individuate the relevant neural symbols and articulate the neural syntax, any mind reading experiment would have to represent the predicted mental content in some natural language, and this translation will be subject to a degree of semantic indeterminacy. A computational system that transforms mentalese into natural language is a kind of translation system, much like a computer program that translates between natural languages. In both cases, we must live with the possibility that different translation systems (whether human or machine) might yield different transla-

[8] On our own view, probabilistic information about the difficulty of guessing is, nevertheless, relevant to evaluating the *proficiency* of a mind reading paradigm. See Section 5.1.

tions, all of which are defensible, but, due to the lack of any inter-subjective standard against which they can be measured objectively, none of which can be shown to trump the others. Notice, moreover, that we cannot choose between translations by asking the experimental subject to referee, because that would be asking for a self-report; an option that we have already set aside. In the absence of a neutral standard against which different labels for mental content can be judged, we suggest that the label generated by the researcher(s) is likely to be as defensible as any other. Consequently, as far as we can see, there is no formulation of the proposal that the label for the mental content must originate in the head of the subject that is compatible with our first anti-convention condition.[9] In light of these two failed attempts to articulate a defensible version of Tong and Pratte's second condition, we doubt that it can be made to work.

We also suspect that the initial plausibility of both conditions Tong and Pratte suggest stems from the intuition that, in order to read a mind, one has to predict something more *mental* than the mundane empirical properties of an experimental setup. This intuition is valuable, but, when expressed in this binary fashion, it is also misleading. In particular, when the intuition is combined with an uncritical attitude toward metaphysics, it fosters the presumption that, first, nature is cleanly divided into mental and non-mental domains, and second, that the distinction between brain reading and mind reading should track that divide. If you think that nature is so divided, then it may be reasonable to hope that some alternative condition, which remains faithful to the spirit of Tong and Pratte's ideas, might yet be discovered. However, if one is already inclined to reject the view that nature is cleanly divided into mental and non-mental properties (as we are), then, not only do the foregoing difficulties become less surprising; they also come to seem more definitive. If a proposed criterion for distinguishing between brain reading and mind reading presumes that nature is cleanly divided into mental and non-mental domains, it is bound to run into similar difficulties.[10]

Tong and Pratte's attempt to distinguish genuine mind reading from mere brain reading is one tactic for setting the explicatory bar high. We now consider an alternative, but nevertheless closely related tactic, which highlights the dis-

[9] The kind of indeterminacy invoked in this argument is the kind that Quine originally described in *Word and Object* (1960), and which he later called "inscrutability of reference." It concerns the choice of natural language label for some snippet of mental content, rather than its cognitive significance.

[10] One version of this point that has recently become popular is the claim that the distinction between mental and non-mental properties is vague. This view has been defended recently by Peter Godfrey-Smith (2020), Michael Tye (2021), and Eric Schwitzgebel (2021).

tinction between decoding experiments that increase scientific understanding, and those that do not.

In a paper from 2015, Kay and Naselaris discuss a limitation of decoding experiments they call representational ambiguity.[11] This is the idea that successful prediction in a decoding experiment does not, on its own, constitute progress toward understanding the function of the cortical region from which data were recorded. To see why, consider the example they provide. Correctly predicting that a participant had been viewing a clip from an action film, rather than one from a romantic comedy, would not suffice to justify the claim that that the cortical region in question has the function of representing film genres. Maybe action films have more visual energy than romantic comedies, and the cortical region is actually dedicated to representing that. Or, maybe the region is dedicated to some other property of visual stimuli that we have yet to think of. Given the data-driven nature of MVPA, accurate prediction of experimental properties does little to constrain neuro-functional hypotheses.

Perhaps this concern about representational ambiguity should lead us to add another condition to our explication of mind reading, according to which mind reading experiments are only genuine if they somehow reflect a deep theoretical understanding of cortical function. We think not. Kay and Naselaris are interested in identifying which experiments will most efficiently improve our understanding of particular brain functions. Our goal is different. It is at least coherent that one might use neural data to learn something about mental content without thereby learning much at all about brain function. In fact, learning about brain function may be more difficult than learning to read minds! This possibility is particularly vivid if we doubt the thesis, implicit in the Kay and Naselaris discussion, that cortical regions are dedicated to the representation of any particular, recognizable, stimulus property. And indeed, there are plenty of reasons to doubt that this kind of stimulus-oriented localizationism is true (Rathkopf, 2013; Anderson, 2014; McCaffrey, 2015).

We have been considering various ways in which the face/house decoding experiment is intuitively underwhelming, and trying to use these intuitions as a guide toward the development of additional conditions we can incorporate into our explication. As it turns out, none of the three conditions suggested thus far have survived scrutiny. So we must now consider other respects in which face/house experiment is underwhelming. In the next section, we con-

[11] Since 2015, similar ideas have been discussed in more detail by both neuroscientists and philosophers, including, most notably Hebart and Baker (2018) and Ritchie et al. (2017).

sider four such respects, and work out what sorts of improvement would bring decoding experiments closer to satisfying our intuitive expectations about the nature of mind reading. However, as we will soon see, those improvements are all inescapably gradual in nature, and their gradualist nature makes them unsuitable material from which to construct necessary conditions. We could insist that genuine mind reading will be accomplished only once some threshold magnitude within each dimension has been reached, but, as we saw in our discussion of Tong and Pratte's second condition, we have no non-arbitrary way of identifying such a threshold.

Moreover, we cannot think of any additional improvements that are not either gradual in nature, or that are otherwise susceptible to the kinds of critical argument we have already given. This fact leads us to the surprising conclusion that the face/house experiment, underwhelming though it may be, *is*, nevertheless, a case of mind reading. We are at least partially justified in regarding it as underwhelming, because it is a maximally simple example of the phenomenon. Experimental design can improve on this starting point in many ways, but no particular improvement will mark a metaphysical transition between decoding, on the one hand, and genuine mind reading, on the other. This may violate our folk psychological intuitions about the meaning of the term "mind reading," but that is a weak reason for skepticism. Our folk psychological intuitions are sometimes misleading, and often betray a tacit acceptance of principles which we would, upon careful reflection, be inclined to reject.

Here then, is our official, categorical explication of mind reading:

> 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.

According to this categorical explication, neuroscientific mind reading is not only possible, it has actually been underway for twenty years. One advantage of this explication is that it avoids tacit commitment to dubious metaphysical principles. One limitation of the explication is that it fails to help articulate the enormous differences between our intuitive judgments about the face/house case, on the one hand, and some of the most cutting edge experiments, on the

other. It is silent about what makes one episode of neuroscientific mind reading more proficient than another.

If our explication of mind reading is to be valuable, it should put us in position to make comparative judgments about real cases. An analogy may help illustrate this point. If we have a clear understanding of the concept of *running*, we should be able to articulate at least a rough method for ranking people with respect to how well they run. We need not expect universal agreement about how to operationalize the ranking (cross-country, track, etc). But if someone were utterly unable to conceive of such an operationalization, we would have reason to doubt that that person understands the concept. Similarly, if we have a clear understanding of the concept of *mind reading*, we should be able to order mind reading experiments with respect to the degree of proficiency they have managed to achieve. In what follows, then, we refine and extend our explication of mind reading by introducing four dimensions along which mind reading might be made more proficient.

5 Dimensions of proficiency

If we compare the most sophisticated decoding experiments being done today with the Haxby et al. experiment from 2001, it is plain that progress has been made. Recent efforts do a better job of fulfilling the folk psychological expectations we impute to experimental work when we choose to frame that work as an attempt to read minds. Until now, those folk psychological expectations have remained latent in the casual judgments of cognitive neuroscientists, and have, therefore, gone largely unnoticed, despite having exerted continuous influence on experimental design. One goal of this section is to identify some of these folk psychological expectations, and to make them more explicit, so that our intuitive standards for the assessment of mind reading experiments are rendered more transparent. Another goal is to reconcile these intuitive standards with the contingencies of scientific practice, which impose constraints on mind reading that are invisible from a purely folk-psychological perspective.

We begin by noting two general constraints, both of which arise from the fact that we are interested in the predictive *capacity* of a decoding model, rather than any particular episode of predictive *performance* (a distinction originally due to Chomsky (1965)). The first general constraint is that we must make a large number of observations. A large sample will help us avoid assigning credit

for lucky guesses, and thereby help us avoid overestimating predictive capacity. The second general constraint is that the predictive task used to generate the data must be sufficiently demanding as to expose the model's predictive limitations. Demanding tasks help us avoid ceiling effects, and thereby help us avoid underestimating predictive capacity.

Before we proceed to our list of four official dimensions, we should comment on the role of predictive accuracy in mind reading. In a sense, it is obvious that, ceteris paribus, the more predictive accuracy a model can achieve, the better. This has led a number of commentators to suggest that accuracy should itself be included as one of the dimensions of mind reading proficiency. We disagree. To see why, it helps to draw attention to the technological character of mind reading experiments, and the fact that technology can change quickly. Our dimensions of proficiency are intended to be useful not only as a way of ordering currently existing technologies, but also as a kind of road map for technological progress in the future. Accuracy cannot serve as a central component in such road map because (i) unlike the other dimensions we are about to introduce, it has a well-defined upper-bound, and (ii) that upper bound has been saturated by some existing experiments. When we try to imagine a future experiment whose accuracy is somehow superior to that of these existing experiments, all we can really do is imagine progress in some *other* dimension. Another way to put the point is that, unlike the four dimensions of proficiency listed below, each of which supports comparisons of the predictive capacities of models *between* experimental designs, predictive accuracy can only sensibly be used as a measure of proficiency *within* a given experimental design. So, although we will mention accuracy frequently, we regard it as too contextually bound to serve as a dimension in its own right.

5.1 Granularity

The intuition behind our first dimension is the natural idea that the boldness of a prediction is inversely proportional to the logical probability of getting it right. Consider an experimental setup in which participants make object category judgments about photographs. A model is trained to label each pattern of neural activation data according to the photograph that triggered it. In one version of this experiment the model is tested on 500 photographs, and achieves near-perfect accuracy. In another, the model is tested on 1000 photographs, and

achieves near-perfect accuracy.[12] In this case, there is a clear sense in which the second experiment has demonstrated a greater capacity for fine-grained mind reading than the first.

In this example, the granularity of the experimental setup is directly proportional to the number of stimuli. This will not generally be the case. Imagine a visual search experiment with natural scene stimuli. In one condition, subjects are told to search for dogs, and in another condition, they are told to search for trees. If individual natural scenes depict both dogs and trees, one stimulus will trigger two distinct patterns of neural activation, depending on the prevailing task instruction. Given this experimental setup, we might attempt to decode the task instruction itself, in addition to the stimulus identity. If successful, we would have an increase in granularity without any corresponding increase in the number of stimuli.[13]

At present, the most fine-grained decoding experiments are probably those designed to reconstruct video stimuli. With thousands of pixels per frame and thousands of frames per clip, these paradigms demand a very large number of predictions. For example, Wen et al. (2017) used a deep convolutional network to extract feature vectors from video frames, and then built a decoder to predict those feature vectors from fMRI data. Finally, they used those predicted feature vectors to reconstruct fuzzy black and white movies that resemble the original video stimuli. From a naive folk-psychological perspective, the idea that we can reconstruct video of something you watched, using only neural data collected while you watched it, is dumbfounding. If that sort of fine-grained reconstruction is possible, one might think, then *any* arbitrarily chosen snippet of mental content can be captured using the same technique.

This wild-eyed assessment is unlikely to be correct. Notice that there is a systematic relationship between the maximum degree of granularity associated with an experimental setup, on the one hand, and the complexity of the thought of the participant, on the other.[14] Imagine that a participant is presented with a picture of a chair. In response, the participant might think: "That's a chair." Or, the participant might think: "That's a Barcelona chair." Or the participant

[12] We say "near-perfect" so as to avoid the concern that a perfect score might indicate a performance ceiling.

[13] This example also helps illustrate why we use the rather broad formulation "properties of the experimental condition," in the categorical explication above.

[14] We have no formal definition of the complexity of thought to offer. We take it to have a relatively intuitive meaning though. In the domain of propositional thought, which can be fairly represented by linguistic expressions, complexity corresponds to the number of inelimnable, non-trivial concepts in a sentence.

might think: "That's a Barcelona chair, designed by Mies van der Rohe in 1929." Given the compositional nature of linguistically structured thought, this sort of complexification exercise can be continued indefinitely.

We can imagine experimental paradigms that evoke the thought "That's a Barcelona chair" but which can only resolve the difference between more coarse-grained options, such "chair" vs. "table." For any level of granularity we may have achieved in a particular experimental setup, a participant might entertain a thought more complex than those our setup permits us to decode, but which is otherwise similar. It seems inevitable, then, that the complexity of thought will always, in principle, be able to outrun the granularity of experimental manipulation. No matter how fine-grained the options, it will be possible to introduce more subtle, more complex thoughts that resist attempts at decoding. This shows that mind reading technology will never be sufficiently powerful to reconstruct any arbitrarily chosen snippet of mental content at a high level of accuracy. It is worth emphasizing the modesty of this conclusion. We do not claim that thought is *typically* more granular than experimental probes are capable of registering. Our claim is only that it is *possible* to entertain thoughts of such complexity. (As anyone who has ever attempted to compose a philosophical essay will be aware, our thoughts are often less complex, less determinate, and less precise than they seem to be, until we manage to articulate them in writing.)

5.2 Generalizability

The intuition behind our second dimension can be illustrated by means of a perceptual learning example. Imagine that a small child sees a Great Dane for the first time, and correctly identifies it as a dog. This untutored sensitivity to the extension of a concept is surely impressive. But it will be more impressive if the child had previously been exposed only to small dog breeds, than if the child had already been familiar with other large ones. In a machine learning context, a similar gradient of proficiency can be recognized. Testing data can be more or less similar to training data. Ceteris paribus, an experiment in which the testing data are radically dissimilar to the training data will generate bolder, more impressive predictions than an experiment in which the training and testing data are nearly identical.

Consider two decoding models that predict equally well, but have been trained differently. The experiment concerns object recognition, and the stimuli

used for testing are photographs of objects presented against a uniformly white background. The first decoding model is trained on photographs of objects that belong to the same object category, also presented against a uniformly white background. The second decoding model is trained on photographs of objects of the same type, but in this case, the objects are located amidst cluttered scenes. In the absence of additional information, the second model should be regarded as the more proficient exemplar of mind reading. Even if the two models perform with equal accuracy, the second model has managed to generate that accuracy despite having had to traverse a greater distance in semantic space between training and testing data.

Cross-subject decoding experiments achieve a form of generalizability that has particular resonance with folk-psychological expectations about what genuine mind reading might involve. In these studies, training data is gathered from one cohort of participants, but tested on another (Haxby et al., 2011). In most cases, all participants are drawn from the same population. In a few recent cases, however, training and testing cohorts are drawn from neurologically distinct populations. For example, van den Hurk et al. (2017) recently decoded auditory stimulus categories from sighted participants after training a model on blind participants. In cases like this, where testing and training cohorts differ systematically, generalizability is pronounced. But in precisely such cases, the decoding model is confronted with systematic functional and anatomical differences between brains, and must, therefore, locate signals in a much noisier informational environment than would otherwise be necessary. In the van den Hurk study, for example, only four stimulus categories were used. The choice to employ this comparatively humble stimulus set reflects a tradeoff between granularity and generalizability. If, as the first half of this paper suggests, mind reading is essentially a data-driven prediction problem, then this tradeoff is inevitable. That inevitability gives us an additional reason to think that progress in mind reading will be made in relatively modest steps, rather than in one qualitative leap, and thereby reinforces the gradualist approach to conceptualizing mind reading that we have been advocating.

Generalizability takes on additional significance when applied to mental content with determinate linguistic structure. As Chomsky (1957) famously stressed, human language is *generative*: given a small store of meaningful words, the combinatorial rules of language allow us generate an unbounded store of meaningful sentences. Moreover, as Adina Roskies (2014) argues, if a decoding model is to capture a wide range of linguistically structured mental content,

it will have to acquire an analogous capacity. Like human children, decoding models are exposed to a relatively small set of meaningful linguistic expressions during training. To become proficient, they will have to learn to decode a much larger set. But decoding the excess content contained in that larger set would be impossible without acquiring some sensitivity to the combinatorial rules of language.

The dimension of generalizability helps to capture the sense in which such sensitivity to the generative properties of language contributes to mind reading proficiency. Consider, for example, a 2018 study by Pereira et al. They describe a linguistic decoding model that takes neural data as input, and delivers a semantic vector as output. Those brain-generated semantic vectors are compared to text-generated semantic vectors, which are high-dimensional word representations based on the frequency with which each word co-occurs with other words in a large text corpus. In one testing condition, participants were presented with novel words not included in the training set. In another testing condition, participants were presented with whole sentences. The decoder achieved considerable predictive success in both conditions. In this work, both testing conditions constitute examples of generalizability, because in both conditions, the testing stimuli differ systematically from the training stimuli. In the second testing condition, however, the difference between training and testing stimuli is distinctively combinatorial. One lesson we can take from this example is that, when a predictive model displays sensitivity to the *generative* properties of language, it is just a special case of *generalizability* in our sense.

5.3 Independence from stimulus

In most of the decoding experiments described thus far, some property of the stimulus is decoded from neural response data. Contrast this with a case in which there are no stimuli at all. In mind-wandering experiments, subjects are told to attend to a fixation cross, and then to think about whatever they please (Chou et al., 2017). Imagine that a participant begins thinking about an essay she is writing on some obscure 17th century painter, and that the model manages to decipher the painter's name from the subject's unconstrained neural activity. That would be a shockingly impressive predictive feat. Why? One reason is that the decoded content was generated by factors unrelated to the design of the experiment. The dimension we call stimulus independence is intended to capture the general principal behind this judgment. We define it as the degree

to which the measured neural response is influenced by factors unrelated to the prevailing experimental condition.

Mind-wandering experiments illustrate an extreme case of stimulus-independent mental activity. Perceptual stimulus identification paradigms exemplify the opposite extreme. There are many interesting cases in the middle. Consider, for example, mental imagery. In some mental imagery experiments, the task is to imagine a stimulus viewed only seconds before the prompt to recall it is given. In other experiments, participants are required to recall a visual stimulus encountered weeks earlier, during a separate experimental session (Horikawa and Kamitani, 2017; Shen et al., 2019). Given the broad range of factors known to influence long-term visual episodic memory, the latter experimental paradigm is more stimulus-independent than the former.

To make progress in the dimension of stimulus-independence, we must confront an interesting epistemological problem. If the experimental setup is not tightly coupled to the mental content we hope to decode, how can a prediction about a property of the experimental condition support faithful discernment of mental content? In other words, how do we know that we have discerned the content correctly? We have no general solution to this problem. We can say, however, that creative experimental design can be used to overcome the problem in particular settings. For example, Rissman et al. (2016) had participants wear cameras around their necks for a period of three weeks, during which time photos of daily activity were taken automatically. Weeks after having taken the cameras off, participants returned to the lab for a brain imaging study in which they were presented with images from their own cameras and images from the cameras of other participants. Rissman et al. achieved nearly perfect accuracy in decoding which pictures corresponded to the experiences of the subject, and which corresponded to experiences of others. Although there is a sense in which neural responses to these photographs are stimulus-dependent, they are not determined by intrinsic features of the photographs. Instead, they are determined by the memory, or lack thereof, evoked by the photograph. This is a clever way to increase stimulus independence while simultaneously ensuring that our judgments of mental content are correct. More generally, we suspect that progress in stimulus-independence of decoding predictions will require novel methods by which data from outside the experiment can be combined with neural data gathered within it.

5.4 Independence from background knowledge

In our categorical explication of mind reading, we said that, in order to count as mind reading, an inference must rely on neural data. There, we made the point by contrasting neural data with behavior, because we wanted to emphasize that the inference should not rely on the quotidian interpretive skills of the researchers. Here, we draw a new contrast between neural data, on the one hand, and background knowledge, on the other.[15]

Imagine you visit a psychic who claims to be able to read minds. At the start of a session she asks you "who is X?" where X is just a very popular name in your area. Assume that you do, in fact, know someone by that name. Does this mean that the psychic has read your mind? Clearly not. The psychic was merely relying on background information. One might suspect that a similar trick is being used in some decoding models. To illustrate this possibility, we introduce another example.

Huth et al. (2016) showed subjects 2 hours of natural movie clips, in which each second of each clip was associated with a small number of hand-coded semantic labels for objects and actions. If an umbrella figured prominently in the scene, one label would be "umbrella." If a car figures prominently in the scene, one label would be "car." On the basis of this scanning, the team trained a machine learning model to associate patterns of neural activity with 1,705 labels. In a separate part of the experiment, subjects were shown movie clips from a distinct set of natural movies. The model would then read the response data and generate labels corresponding to the content of the movie a subject had been watching.

The 1,705 labels were taken from a standard semantic database for commonly used words. However, this database allowed for labels with overlapping meaning. For example, one label might be "car" and another label might be "station wagon." If the decoding model draws decision boundaries for these labels separately, it might conclude that the probability that a given pattern of brain activity was prompted by a "station wagon" is higher than the probability that it was prompted by a "car." To avoid such logical impossibilities, Huth et al. built a model that explicitly limited the conditional probabilities of conceptually related stimuli. The conditional probabilities were drawn from WordNet,

[15]To preserve the distinction between this dimension and the dimension of generalizability, we emphasize that training data do not count as background knowledge. Here, we are thinking instead of various ways in which researcher priors regarding the predictive target can be used to tune parameters by hand, rather than by machine learning.

a well-known lexical database. These conditional probabilities constitute a kind of background knowledge, and they are crucial to the predictive success of the model.

Does the top-down imposition of these conditional probabilities tell against the degree of mind reading proficiency achieved by the Huth et al. model? One might think not. After all, if you read a book in which cars and station wagons figure prominently, your background knowledge that a station wagon is a type of car contributes to your ability to understand the story. If such background knowledge positively contributes to one's competence in reading books, why consider it a threat to one's competence in reading minds? To answer this question, consider a variation on the Huth et al. model in which the role of background knowledge is exaggerated. In this variation, a list of cars that appear in the video collection is explicitly encoded at the outset, along with the color of each car. Such information would massively simplify the task of predicting car color in any given clip. Under these conditions, correctly predicting that the car in the film is blue, say, is driven largely by background knowledge rather than by neural data. In this exaggerated example, the role of background knowledge in generating the prediction is relevantly like the role of background knowledge in the psychic's "Who is X?" trick. Although you cannot eliminate background knowledge, this example illustrates that some experimental designs rely more heavily on background knowledge than others. If we could build a model that managed to achieve the same degree of predictive success without relying on car color data, it would deserve to be counted as a more proficient exemplar of mind reading technology.

Our example also illustrates, once again, that when trying to read minds, tradeoffs are inevitable. In this case, background knowledge facilitates progress in the granularity dimension (as indicated by the 1,705 labels used), but only by dragging down the aggregate degree of mind reading proficiency. This provides an additional reason to believe that progress will be a matter of degree. Each progressive step will be both valuable and interesting, but none is likely to mark a metaphysical distinction between brain reading and mind reading, in the way Tong and Pratte have suggested.

6 Conclusion

In the explication above, we have attempted to articulate what kind of achievement neuroscientific mind reading might be, what kind of evidence is relevant to determining whether it has been achieved, and how to compare mind reading episodes with respect to their degree of proficiency. On the view we have developed, 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. One implication of this view is that neuroscientific mind reading has, in fact, already been accomplished. In light of that result, the comparative aspect of our explication becomes particularly important because it helps accommodate conflicting folk-psychological intuitions about what mind reading is supposed to be.

The comparative analysis gave rise to four dimensions along which proficiency can be assessed. It also showed that experimental design forces us to confront tradeoffs between these dimensions. Each experimental setup must be carefully tuned toward the particular kind of mental content one hopes to predict. As a result, we expect progress in neuroscientific mind reading to be thoroughly piecemeal.

The picture we have developed here contrasts with the picture of mind reading entertained in popular imagination in at least two ways. First, contrary to popular imagination, neuroscientific mind reading is not the sort of achievement that is unlocked by any particular technological development. It is better described as a heterogeneous family of experimental design and analysis ideas. There is no core technology which, once developed, leads inexorably toward better decoding capabilities. Additional progress will require creative scientific thinking at every step.

The second contrast with popular imagination concerns the source of fascination with neuroscientific mind reading that we discussed at the outset. There, we said that the prospect of neuroscientific mind reading is fascinating in large part because it enables a mode of access to the thoughts of other people that does not rely on conventional, public symbols. When we emphasize the possibility of using technology to circumvent reliance on conventional, public symbols, it becomes tempting to think that the technology in question will offer a more direct, less mediated form of communication than conventional public forms of

communication typically allow. On the picture we have developed, this tempting inference should be avoided. Despite the fact that neuroscientific mind reading does circumvent reliance on conventional public symbols in a way that is novel, it is, nevertheless a rather indirect method of discerning mental content. It is indirect in the sense that decoding any particular snippet of mental content will require a carefully curated and typically elaborate experimental design. The effort required to discern mental content from neural data is, and is destined to remain, greater than that required by more traditional means of communication.

When we try to imagine a future in which neuroscientific mind reading has advanced considerably, it is easy to be mislead by folk-psychological intuitions about how minds work, and, as a result, to end up with a picture that has little to do with empirical reality. Here, we have tried to offer an alternative, empirically grounded picture of what neuroscientific mind reading is, and what it might become. To conclude, we would like to stress that, although our arguments are designed to have a sobering effect, they are not designed to have a deflationary one. We do not wish to diminish the sense of fascination and excitement evoked by the possibility of discerning mental content directly from neural activity. Instead, we hope that by placing empirical constraints on our imaginative efforts, we put ourselves in better position to understand how neuroscientific mind reading techniques can contribute to the larger scientific enterprise of figuring out how the mind works.

References

Michael L Anderson. *After phrenology.* MIT Press, 2014.

Rudolf Carnap. *Logical foundations of probability.* University of Chicago press, 1962.

Noam Chomsky. *Syntactic structures.* Walter de Gruyter, 1957.

Noam Chomsky. *Aspects of the theory of syntax.* MIT Press, 1965.

I-han Chou. Telepathy? I think not. *The Action Potential Blog*, February 2012. URL `http://blogs.nature.com/actionpotential/2012/02/telepathy-i-think-not.html`.

Ying-hui Chou, Mark Sundman, Heather E Whitson, Pooja Gaur, Mei-Lan Chu, Carol P Weingarten, David J Madden, Lihong Wang, Imke Kirste, Marc Joliot, et al. Maintenance and representation of mind wandering during resting-state fmri. *Scientific reports*, 7:40722, 2017.

Jerry A Fodor. *The language of thought.* Harvard university press, 1975.

Peter Godfrey-Smith. *Metazoa: Animal minds and the birth of consciousness.* Harper Collins UK, 2020.

Ryohei P Hasegawa, Yukako T Hasegawa, and Mark A Segraves. Neural mind reading of multi-dimensional decisions by monkey mid-brain activity. *Neural Networks*, 22(9):1247–1256, 2009.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

Martin N Hebart and Chris I Baker. Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, 180:4–18, 2018.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8:15037 EP –, 05 2017. URL https://doi.org/10.1038/ncomms15037.

Virginia Hughes. "Reading minds" with fMRI, October 2011. URL https://www.lastwordonnothing.com/2011/10/13/reading-minds-with-fmri/.

Alexander G Huth, Tyler Lee, Shinji Nishimoto, Natalia Y Bilenko, An T Vu, and Jack L Gallant. Decoding the semantic content of natural movies from human brain activity. *Frontiers in systems neuroscience*, 10:81, 2016.

Christopher Intagliata. Neuroscientists take an important step towards mind reading, 2008. URL https://www.scientificamerican.com/article/can-you-read-my-mind/.

Arnav Kapur, Shreyas Kapur, and Pattie Maes. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*, pages 43–53. ACM, 2018.

Joseph B McCaffrey. The brain's heterogeneous functional landscape. *Philosophy of Science*, 82(5):1010–1022, 2015.

Kenneth A. Norman, Sean M. Polyn, Greg J. Detre, and James V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, 10(9):424 – 430, 2006. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2006.07.005. URL http://www.sciencedirect.com/science/article/pii/S1364661306001847.

Matt Peckham. Scientists can (almost) read your mind, turn thoughts into movies, September 2011. URL http://techland.time.com/2011/09/23/scientists-can-almost-read-your-mind-turn-thoughts-into-movies/.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, 2018. doi: 10.1038/s41467-018-03068-4. URL https://doi.org/10.1038/s41467-018-03068-4.

Russell A Poldrack. *The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts*. Princeton University Press, 2018.

Willard Van Orman Quine. *Word and object*. MIT Press, 1960.

Charles Rathkopf. Neural reuse and the nature of evolutionary constraints. In Fabrizio Calzavarini and Marco Viola, editors, *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience*, volume 17 of *Studies in Brain and Mind*. Springer Nature, 2021.

Charles A Rathkopf. Localization and intrinsic function. *Philosophy of Science*, 80(1):1–21, 2013.

Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2):818 – 825, 2010. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2009.11.084. URL `http://www.sciencedirect.com/science/article/pii/S1053811909012701`.

Jesse Rissman, Tiffany E Chow, Nicco Reggente, and Anthony D Wagner. Decoding fmri signatures of real-world autobiographical memory retrieval. *Journal of cognitive neuroscience*, 28(4):604–620, 2016.

J Brendan Ritchie, David Michael Kaplan, and Colin Klein. Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, 70(2):581–607, 2017.

Pieter R Roelfsema, Damiaan Denys, and P Christiaan Klink. Mind reading and writing: the future of neurotechnology. *Trends in cognitive sciences*, 22 (7):598–610, 2018.

Adina Roskies. Mindreading and privacy. In Michael Gazzaniga and George R. Mangun, editors, *The Cognitive Neurosciences*, chapter 85. MIT Press, 2014.

Eric Schwitzgebel. Borderline consciousness, when it's neither determinately true nor determinately false that experience is present. Online talk at Kinds of Intelligence 3, Cambridge University, September 2021.

Rebecca Seligman, Suparna Choudhury, and Laurence J Kirmayer. Locating culture in the brain and in the world. In *The Oxford Handbook of Cultural Neuroscience*. Oxford University Press, 2016.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

Frank Tong and Michael S Pratte. Decoding patterns of human brain activity. *Annual review of psychology*, 63:483–509, 2012.

Michael Tye. *Vagueness and the Evolution of Consciousness: Through the Looking Glass*. Oxford University Press, 2021.

Job van den Hurk, Marc Van Baelen, and Hans P Op de Beeck. Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22):E4501–E4510, 2017.

Gretchen Vogel. Research on communication with completely paralyzed patients prompts misconduct investigation, Apr 2019. URL `https://www.sciencemag.org/news/2019/04/research-communication-completely-paralyzed-patients-prompts-misconduct-investigation`.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, 2017.

Clare Wilson. Mind-reading implant can decode what your ears are hearing, January 2019. URL `https://www.newscientist.com/article/2192116-mind-reading-implant-can-decode-what-your-ears-are-hearing/`.