Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/63/2024.17941

Reverse Inference: Decoding of Brain Activity and Cognitive Process

Zirui Huang

Cognitive Systems Program, University of British Columbia, University Endowmenland, Canada

ziruihh@student.ubc.ca

Abstract. Cognitive neuroscientists rely on functional neuroimaging techniques and behavioral assays to investigate correlation between brain activation and cognitive processes. Researchers would also infer what cognitive process is engaged under a condition based on neuroimaging data. This is referred to as reverse inference or encoding paradigm and it has generated longstanding discussions among cognitive neuroscientists, data scientists and philosophers. The consensus is that it should be done with rigorous statistical methods and careful interpretations. Statistical methods and large collaborative databases and data processing tools have been developed to build classificatory or predictive models of reverse inference. However, these tools are not without pitfalls. The problem of cognitive process further complicates the problem since it is a latent variable and the wide discrepancy on the taxonomy and ontology of cognitive processes within the field. Online collaborative project has been set up to combat this problem, while others try to circumvent pre-established concepts by extracting latent variables from existing data or by focusing on the evolutionary prerequisite of cognition. A fundamental limitation of reverse inference is its dependency on correlational data, which lacks the explanatory power interventionist studies have. Models that seek to establish causal relation from time-series data can only provide weak inferences. Furthermore, the dynamic complexity of the brain network complicates the mechanism. This review provides an overview of reverse inference in cognitive neuroscience. It discusses advances in methods, limitations, and conceptual issues inherent within reverse inference, particularly addressing the challenges mentioned above.

Keywords: Reverse Inference. Cognitive Neuroscience. Brain decoding.

1. Introduction

Scientific reasoning, unlike pure deductive or mathematical reasoning, relies on making risky inference based on empirical data. In another word, scientific statements should be falsifiable. The study of cognitive neuroscience also relies on making risky inferences. Robust statistical tools, tightly controlled task experiments and high precision imaging instruments have been developed to refine inferences in cognitive neuroscience research. However, even with advanced methodologies, scientific inference should still be made with caution with careful interpretation of empirical data. Reverse inference is a central issue that exposes the current limitation and pitfalls in cognitive neuroscience research where the fundamental problems does not only lie in the technology researchers have at their disposal, but the abilities of the researchers to be aware of the deeply entrenched methodological limitations. A lack of

DOI: 10.54254/2753-8818/63/2024.17941

understanding and awareness of methodological limitation can greatly impair researchers' judgment and mislead them to make unwarranted claims.

Reverse inference, in short, is the process from inferring mental states, cognitive processes or any innate and latent psychological processes from measured brain activities. Poldrack's article on this topic in 2006 [1] generated lasting discussion on this topic and its trailing implications. It has inspired the fields to develop statistical tools to support more rigorous methods to help researchers making sound claims [2-11]. As cognitive neuroscience increasingly relies on complex models and deep learning techniques to analyze brain data, the interpretation of these results becomes more intricate, raising important questions about the validity and reliability of reverse inference as a tool for understanding human cognition [11, 12]. Statistical analysis, however, is not the sole issue underlying reverse inference. In the existing literature, there is much heterogeneity among the definitions and operationalization of concepts of different cognitive processes and their relationships [1,7,13]. Furthermore, lack of interventionist data in cognitive neuroscience exacerbates the limitation of current use of reverse inference, such as lack of explainability [14-16].

This article delves into the logic, goals, and inherent challenges of reverse inference within cognitive neuroscience, exploring the nuances of statistical methodologies, the ongoing debates surrounding cognitive ontology, and the complex relationship between observation, perturbation, prediction, and causation in this rapidly evolving field.

2. Logic and Goal of Reverse Inference in the Context of Cognitive Neuroscience

The archetypal cognitive neuroscience experiment goes as follows: certain behavioral tasks are randomized as the independent variable. Some types of behavioral response (i.e. eye-tracking, response time, etc.) are measured along with brain activities with hemodynamic measure (e.g. Functional Magnetic Resonance Imaging (fMRI) or electrophysiology (e.g. Electroencephalography (EEG) [1,17]. Putatively, a certain cognitive process is elicited by the behavioral task and the behavioral response is measured to confirm whether this cognitive process is activated or the level of activation. The measured data of brain activities is then analyzed along with the behavioral outcome to find correlation between brain activation and engagement of cognitive processes of interest [17]. The identification of correlation between brain activation and engagement of cognitive processes is considered as forward inference [17]. Note the casual chain in this study design: the only manipulation is the randomization of behavioral manipulation. Causal claims can be made between behavioral task and brain activities and between task and behavioral outcome. No causal claims can be made between the brain activation and behavioral outcome (by extension, the engagement of cognitive process).

Reverse inference in cognitive neuroscience is the act of inferring cognitive process engagement by brain activation data [1]. This is also referred to as the "decoding model" [10]. Inference of this type requires prior literature to establish correlation between cognitive process and neural activities, particularly forward inference studies. Reverse inference is often characterized as an abductive reasoning in the sense that it is an "inference to the best explanation" where the most plausible conclusion is made based on available but incomplete evidence [1,18-24]. In probabilistic terms, it is to determine the conditional probability of the involvement of a particular cognitive process given that a specific neural activity is observed in an individual, specify it as P(COG|NEU). Bayes' formula could be used to calculate this probability, the form of

$$P(COG|NEU) = \frac{P(NEU|COG)\times P(COG)}{P(NEU)} = \frac{P(NEU|COG)\times P(COG)}{P(NEU|COG)\times P(\sim COG)+P(NEU|\sim COG)\times P(\sim COG)} \quad (1)$$

P(COG) would be the prior probability of whether the cognitive process of interest is elicited in general, as known as the base rate, similarly P(NEU) would be the base rate of the specific brain activity under investigation. "~" denotes negation, thus P(NEU|~COG) denotes the conditional probability of the presence of a specific neural activity given that the cognitive process of interest is absent. The posterior probability calculated or the likelihood ratio (P(COG|NEU)/P(NEU), also known as Bayes factor) could be used to determine the selectivity between the cognitive process and brain activity [1,19-

DOI: 10.54254/2753-8818/63/2024.17941

24]. The level of selectivity, or the strength of correspondence between brain activity and cognitive process are the criterion for acceptable reverse inference [1]. Strong correlation is not enough since a brain activity could be observed under a variety of cognitive processes engagement. Bayes' theorem provides the tool to evaluate that specificity between these two variables. Brain activities measurement here could be either a local activation (structural location) or a pattern-based response (global or local) [2,3,20]. In the case of pattern-based response, it could be the pattern of multiple voxels' activation at a particular region, the functional connectivity of a brain network, or the wave pattern recorded by the EEG/MEG. Reverse inference can also work in the other direction, by inferring how the brain is activated given a cognitive process is active.

3. Statistics of Reverse Inference and Its Challenges

Cognitive neuroscience relies on human participants for the study of higher-level cognition such as language processing and production, thus most studies use noninvasive neuroimaging techniques to capture neural activities. As mentioned above in section 1, this approach generates correlational data between brain activities and cognitive processes engagements. To draw conclusions from correlational data, statistical analysis is indispensable.

Statistical and consolidated databases have been developed to facilitate formal and vigorous reverse inference. However, as statistical vigor increases, the data processing pipeline becomes more complicated as well. Multivariate pattern analysis (MVPA) is a common choice among researchers for pattern classification [20]. The extracted pattern can then be used to predict whether a cognitive process is active or not based on brain activation.

Deep learning methods are now common for this purpose, where input data are passed through multiple layers of nonlinear functions in order to reach an abstract representation of the dataset [8,12]. However, the pipeline of data processing with deep learning is difficult to interpret how the model makes decisions regarding decoding. Certain techniques have been developed to address this issue such as sensitivity analyses, background decompositions and reference-based attribution [11,12]. All of the above techniques aim to identify relevant features that deep learning models used to make predictions, either by analyzing each layer backward or by using a reference feature as comparison.

The predictive power of deep learning is impressive, but it is not immune to methodological issues. In fact, because of its complex data pipeline, it is more susceptible to having methodological issues that are masked by its apparent superior performance. One problem is hidden stratification [12], where trained deep learning models can perform well on average within a given dataset but have difficulties discriminating within subclasses. Additional stratification of the dataset before training, either automated or manually, could help combat hidden stratifications. Deep learning models are also overly sensitive to confounds because of its strong ability to extract patterns [12]. It could learn to make decisions based on features that should not be considered. For example, a deep learning model trained on neuroimaging data might be sensitive to specific noise due to head motions (as in fMRI). To test for robustness and potential confounds, the trained model should be cross-validated with datasets the model was not trained with.

4. Problems with Cognitive Ontology

Relating established psychological constructs in the literature on neural architecture and pattern of activities is a common top-down approach, but the problem of cognitive process is that it is a latent variable [1,3,13]. Cognitive processes' presence cannot be directly observed but only indirectly inferred. Behavioral tasks only putatively recruit cognitive processes into action. Often, researchers would conflate the task with the cognitive process itself. It is difficult to isolate cognitive processes even under behavioral interventions. These constructs are also subjects of fraught debates among psychologists. Although experimental protocols have been in use for many studies that take them for granted, academics cannot reach a satisfactory consensus that assigns them with stable concepts. Cognitive Atlas is an open collaborative project founded by Poldrack [7] that aims to develop a coherent and

DOI: 10.54254/2753-8818/63/2024.17941

comprehensive cognitive taxonomy that can represent the consensus among researchers. It has been utilized in some studies but not widespread.

One of the main motivations of cognitive neuroscience is to enrich our understanding of human's cognitive architecture from a neurobiological perspective. Neuroimaging data can be used to discern whether a psychological construct should be considered with ontological status. Lenartowicz et al. [25] have used machine learning to classify different neural activation patterns under different tasks that purportedly operationalize different cognitive processes such as working memory, response selection, response inhibition, and task switching. The classifiers can accurately differentiate activation under different behavioral conditions but task switching, thus Lenartowicz et al. concludes that task switching should not be considered as a discrete cognitive process. Similarly, a data-driven approach also harnesses the vast amount of neuroimaging data under different behavioral conditions to identify latent variables that link neural activation patterns and task conditions [26]. The latent variables identified are assumed as basic units of cognition. By suspending judgment on cognitive ontology, this approach has the advantage of not relying on potentially unreliable cognitive theories that could bias the result and uncovering novel cognitive processes.

Another approach called phylogenetic refinement draws upon an evolutionary perspective, viewing cognitive processes as adaptive strategies in face of evolutionary pressure [27]. Unlike evolutionary psychology, which focuses on sexual selection and social cognition, phylogenetic refinement focuses on brain evolution. Similarly, this approach deliberately circumvents problems of a top-down pre-defined cognitive ontology which relies on folk-psychology to avoid biases in empirical investigation. As the vertebrate nervous system evolves, its repertoire of behavior bifurcates and sophisticates from its previous system. It is important to understand higher-level behaviors (such as cognitive processes) in terms of how it arises from basic behaviors. This perspective deflates cognition as behavior, but affirming that behavior as innately originated where behavioral output is not triggered by input, but the inputs modulate the system's output and its output controls the inputs. It reframes the question of cognition as how it arises within the brain of which has evolved to have an interactive behavior repertoire.

5. Observation, Perturbation, Prediction and Causation

It is important to remember that the decoding paradigm utilizes correlational data. The feature extracted by deep learning models to predict cognitive processes might only be epiphenomenal. The success of a trained decoding model in accurately predicting cognitive process engagement does not mean that it can identify the brain activities pattern necessarily for that cognitive process [28]. Brain stimulation studies (intracranial or noninvasive) are usually done with patient populations, which compromises their ecological validity as it is difficult to generalize their result as an explanatory account of neural mechanisms underlying cognition.

In fields such as economics and public health, researchers would infer causation from observational data. In neuroscience, Granger causality analysis (GCA) and Dynamic Causal Model (DCM) [16,28] has been used to identify interaction among brain regions with fMRI data. However, a model using correlational data to infer causality will require substantial assumptions about the data and methodology which cannot always be satisfied. GCA necessitates time series data over neural variables at a consistent sampling rate. Any mismatch, especially in fMRI measurements, can lead to erroneous conclusions about causal relationships due to undersampling. To yield reliable causal knowledge with GCA, substantial background knowledge is essential. Without this, GCA may not effectively distinguish between correlation and causation, particularly in complex neuroimaging contexts. Also, the temporal structure of the data is crucial. GCA can fail if there is a mismatch between the actual timescale of neural activity and the measurement timescale, which is common in fMRI studies.

The human nervous system is also depicted as a dynamic complex control system which is characterized by its nonlinearity, and feedback-loops [29,30]. Causation can be interpreted two-fold. The classical view of causation assumes a mechanism with separable units and unidirectional information flow. However, the brain has network architecture and functions as a nonlinear dynamical system where causality is topological. For example, between brain regions interactions are often

DOI: 10.54254/2753-8818/63/2024.17941

bidirectional and cyclical [31]. Pessoa and György [29,30] have expressed skepticism on the concept of causation in neuroscience. Even under strict and localized manipulation (such as a single neuron), the resulting changes can lead to secondary effects that are not directly related to the initial perturbation. This makes it difficult to ascertain whether observed outcomes are due to the direct manipulation or to these secondary changes in the network. Furthermore, neurons are part of intricate circuits, and perturbing one part of the network can unpredictably affect other areas [28]. This interconnectedness can obscure the true causal pathways, as the network may respond in unexpected ways. In complex systems, causation can be reciprocal or circular, meaning that the relationships between variables are not linear. This complicates the interpretation of results, as the cause may not be easily identifiable.

6. Conclusions

This review delineates the basics of reasoning of reverse inference in cognitive neuroscience research, the concerns of this mode of inquiry, related conceptual issues such as cognitive ontology and causation, the relevant statistical methodology and their limitations, and proposed solutions for improvement. The field of cognitive neuroscience has significantly advanced thanks to the tools developed with data science, machine learning and neuroimaging techniques. Motivation for such advancement is partially due to researcher's concerns of limitations of reverse inference. However, as discussed above, reverse inference problems extend beyond data analysis. Conceptual issues are as pertinent as methodological issues, if not more. Often, conceptual issues constrain what methodology is allowed.

As methodology advances and sophisticates, more possibilities and opportunities for generating knowledge emerge. At the same time, some unsolved fundamental limitations will emerge as new pitfalls and challenges. This review hopes to guide current researchers of cognitive neuroscience to navigate these challenges.

References

[1] Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences, 10(2), 59–63.

[2] Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature Methods, 8(8), 665–670.

[3] Poldrack, R. A. (2011). Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. Neuron, 72(5), 692–697.

[4] Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. NeuroImage, 62(2), 852–855.

[5] Mather, M., Cacioppo, J. T., & Kanwisher, N. (2013). How fMRI Can Inform Cognitive Theories. Perspectives on Psychological Science, 8(1), 108–113.

[6] White, C. N., & Poldrack, R. A. (2013). Using fMRI to Constrain Theories of Cognition. Perspectives on Psychological Science, 8(1), 79–83.

[7] Poldrack, R. A., & Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. Annual Review of Psychology, 67(1), 587–612.

[8] Wang, X., Liang, X., Jiang, Z., Nguchu, B. A., Zhou, Y., Wang, Y., Wang, H., Li, Y., Zhu, Y., Wu, F., Gao, J., & Qiu, B. (2020). Decoding and mapping task states of the human brain via deep learning. Human Brain Mapping, 41(6), 1505–1519.

[9] Zhuo, C., Li, G., Lin, X., Jiang, D., Xu, Y., Tian, H., Wang, W., & Song, X. (2021). Strategies to solve the reverse inference fallacy in future MRI studies of schizophrenia: A review. Brain Imaging and Behavior, 15(2), 1115–1133.

[10] Menuet, R., Meudec, R., Dockès, J., Varoquaux, G., & Thirion, B. (2022). Comprehensive decoding mental processes from Web repositories of functional brain images. Scientific Reports, 12(1), 7050.

[11] Thomas, A. W., Ré, C., & Poldrack, R. A. (2023). Benchmarking explanation methods for mental state decoding with deep learning models. NeuroImage, 273, 120109.

DOI: 10.54254/2753-8818/63/2024.17941

[12] Thomas, A. W., Ré, C., & Poldrack, R. A. (2022). Interpreting mental state decoding with deep learning models. Trends in Cognitive Sciences, 26(11), 972–986.
[13] McCaffrey, J., & Wright, J. (2022). Neuroscience and Cognitive Ontology: A Case for Pluralism. In F. De Brigard & W. Sinnott-Armstrong (Eds.), Neuroscience and Philosophy (pp. 427–466). The MIT Press.
[14] Bergmann, T. O., & Hartwigsen, G. (2021). Inferring Causality from Noninvasive Brain Stimulation in Cognitive Neuroscience. Journal of Cognitive Neuroscience, 33(2), 195–225.
[15] Weichwald, S., & Peters, J. (2021). Causality in Cognitive Neuroscience: Concepts, Challenges, and Distributional Robustness. Journal of Cognitive Neuroscience, 33(2), 226–247.
[16] Siddiqi, S. H., Kording, K. P., Parvizi, J., & Fox, M. D. (2022). Causal mapping of human brain function. Nature Reviews Neuroscience, 23(6), 361–375.
[17] Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. Trends in Cognitive Sciences, 10(2), 64–69.
[18] Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. NeuroImage, 84, 1061–1069.
[19] Glymour, C., & Hanson, C. (2016). Reverse Inference in Neuropsychology. The British Journal for the Philosophy of Science, 67(4), 1139–1153.
[20] Del Pinal, G., & Nathan, M. J. (2017). Two Kinds of Reverse Inference in Cognitive Neuroscience. In The Human Sciences after the Decade of the Brain (pp. 121–139). Elsevier.
[21] Krueger, J. I. (2017). Reverse Inference. In S. O. Lilienfeld & I. D. Waldman (Eds.), Psychological Science Under Scrutiny (1st ed., pp. 108–122). Wiley.
[22] Nathan, M. J., & Del Pinal, G. (2017). The Future of Cognitive Neuroscience? Reverse Inference in Focus. Philosophy Compass, 12(7), e12427.
[23] Calzavarini, F., & Cevolani, G. (2022). Abductive reasoning in cognitive neuroscience: Weak and strong reverse inference. Synthese, 200(2), 70.
[24] Coraci, D., Calzavarini, F., & Cevolani, G. (2023). Reverse Inference, Abduction, and Probability in Cognitive Neuroscience. In L. Magnani (Ed.), Handbook of Abductive Cognition (pp. 1523–1549). Springer International Publishing.
[25] Lenartowicz, A., Kalar, D. J., Congdon, E., & Poldrack, R. A. (2010). Towards an Ontology of Cognitive Control. Topics in Cognitive Science, 2(4), 678–692.
[26] Yeo, B. T. T., Krienen, F. M., Eickhoff, S. B., Yaakub, S. N., Fox, P. T., Buckner, R. L., Asplund, C. L., & Chee, M. W. L. (2015). Functional Specialization and Flexibility in Human Association Cortex. Cerebral Cortex, 25(10), 3654–3672.
[27] Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. Attention, Perception, & Psychophysics, 81(7), 2265–2287.
[28] Danks, D., & Davis, I. (2023). Causal inference in cognitive neuroscience. WIREs Cognitive Science, 14(5), e1650.
[29] Buzsáki, G. (2019). Causation and Logic in Neuroscience. In G. Buzsáki, The Brain from Inside Out (1st ed., pp. 33–52). Oxford University PressNew York.
[30] Pessoa, L. (2023). The Entangled Brain. Journal of Cognitive Neuroscience, 35(3), 349–360.
[31] Harnack, D., Laminski, E., Schünemann, M., & Pawelzik, K. R. (2017). Topological Causality in Dynamical Systems. Physical Review Letters, 119(9), 098301.