On the Logic of Measuring Neural Correlates of Consciousness

Johannes Kleiner[1,2,3,4]

[1]Institute for Psychology, University of Bamberg
[2]Munich Center for Mathematical Philosophy, LMU Munich
[3]Graduate School of Systemic Neurosciences, LMU Munich
[4]Association for Mathematical Consciousness Science

Abstract. This paper presents a mathematical analysis of the logic of measuring Neural Correlates of Consciousness (NCCs). Starting from the canonical definition of NCCs provided by Crick and Koch (1990) and Chalmers (2000), a series of lemmas and theorems are provided which show how NCCs can be discovered if empirical data about the co-activation of neural states and states of consciousness is available. The result of this analysis is a new method to measure NCCs, which we preliminarily call Co-Activation Analysis (CoAA), that might complement or extend existing methods such as contrastive analysis and decoding. CoAA might be of interest for experiments because it does not require data from near-threshold conditions, is compatible with all major conceptions of states of consciousness (including, e.g., micro-phenomenological notions and global states of consciousness), can be applied to most conceptions of neural or computational states (including, e.g., those provided by Predictive Processing/Active Inference, or those attained in more ecological conditions), and helps alleviate the problem of confounders. Furthermore, we show as a theorem that as far as the logic of measurement is concerned, if applied to data from contrastive analysis studies, CoAA improves upon the result provided by contrastive analysis. The paper is a purely theoretical contribution. It is presented in the hope that the mathematics developed here can support and improve the empirical search for NCCs.

Contents

1. Introduction

The search for Neural Correlates of Consciousness (NCCs) is an important pillar of consciousness science. It carries the hope of finding constraints on how conscious experiences and brain-states relate that do not rely on theories of consciousness. Because of their theory-independence, NCCs may play an important role in theory-selection and theory-construction in consciousness science, and may offer important empirical guiderails for the further development of the field.

The canonical method used to measure NCCs to date is contrastive analysis (Baars, 1986). Contrastive analysis presumes a binary distinction in conscious experiences—whether a subject experiences a particular stimulus or content consciously, for example—and aims at measuring the difference (contrast) in neural activity between two conditions that only differ in the two states of consciousness.

The contrastive methodology has enabled impressive empirical progress, cf. (Lepauvre & Melloni, 2021; Förster, Koivisto, & Revonsuo, 2020; Koch, Massimini, Boly, & Tononi, 2016; Singer, 2014; Tononi & Koch, 2008; Metzinger, 2000), but it has also been subjected to thorough criticism, identifying a number of challenges for a successful empirical identification of the NCC proper that are not easy to meet, cf. (Overgaard, 2004; Hohwy, 2009; Aru, Bachmann, Singer, & Melloni, 2012; Peters, Kentridge, Phillips, & Block, 2017; Mudrik, Hirschhorn, & Korisky, 2024). Most notably, it is "necessary to control for all of [the] potential (known) confounds in order to look for the neural activity that systematically relates to consciousness" (Lepauvre & Melloni, 2021, p. 10). Such confounds include processes related to reporting, working memory, decision making, prior expectations, and particular task demands.

The empirical challenges in measuring NCCs have motivated explorations of better measurement strategies, for example in (Lau & Passingham, 2006; Frässle, Sommer, Jansen, Naber, & Einhäuser, 2014; Matthews, Schröder, Kaunitz, Van Boxtel, & Tsuchiya, 2018; Matthews et al., 2018; Nani et al., 2019; Cohen, Ortego, Kyroudis, & Pitts, 2020; Dellert et al., 2021; Demertzi, 2023), of the decoding approach to NCCs (Haynes, 2009), of conceptual re-assessments of what an NCC should be taken to be in the first place, for example in (Fink, 2016; Hohwy & Seth, 2020; Klein, Hohwy, & Bayne, 2020; Kriegel, 2020; Fink, Kob, & Lyre, 2021; Wiese & Friston, 2021; Paßler, 2023), and reflections on measures of consciousness, for example in (Fleming & Lau, 2014; Overgaard, 2015; Michel, 2019; Mazor & Fleming, 2020; Marvan & Polák, 2020; Pauen & Haynes, 2021; Michel, 2023).

The present paper investigates the logic of measuring NCCs. Starting from the definition of NCCs that is broadly accepted in the field—the 'canonical' definition provided by Crick and Koch (1990) and Chalmers (2000)—and presuming data about the co-activation of neural states and states of consciousness, it aims to identify the logical connections that afford for an NCC to be empirically discovered.

The result of this investigation, surprisingly, provides an alternative approach to measure NCCs that is different from contrastive analysis and decoding. We found that this *Co-Activation Analysis*, as we preliminarily call it, does not require data from near-threshold conditions (Section 10), is compatible with all major conceptions of states of consciousness (Section 10), can be applied to many meaningful conceptions of neural systems (Section 9), and may help to appease or resolve the problem of confounders (Section 8). Furthermore, it can be shown to reproduce the result of contrastive analysis to the extent of the latter's power (Section 6, Theorem 19).

In light of its aim, it may not be surprising that this paper is a purely theoretical contribution. No data has been harmed in its conception. Accordingly, it goes without saying that the results presented here constitute, if anything, only first steps in the development of a full-fledged empirical programme. While the mathematical side of the proposal has been carefully checked, empirical application might bring refinements or potentially revisions.

The paper is structured as follows. In Section 2, we introduce the canonical definition of

NCCs, which embodies the default understanding of NCCs in the consciousness science community. In Section 3, we explain which sort of empirical data is presupposed by the analysis presented here (data about the co-activation of neural states and states of consciousness). In Section 4 we explain how the analysis of the logic of measuring NCCs proceeds. The analysis is then given in Section 5, running through the constituents of the canonical definition of NCCs one by one. Section 6 is devoted to the comparison with contrastive analysis, and Section 7 explains how statistical testing can be applied to the methodology developed here. Section 8 discusses the problem of confounders. Section 9 explains how the methodology proposed here can be used to search for Computational Correlates of Consciousness and similar notions. Section 10 discusses examples of states of consciousness that the analysis can be applied to, and Section 11 summarizes the resulting picture and explains how the methodology would be applied in practice, including a preliminary assessment of the possibility of reanalysing data from contrastive analysis studies with the co-activation methodology. A conclusion and short outlook on possible future work is offered in Section 12.

The following analysis of the logic of measurement does not apply to the decoding approach to NCCs. This is because the decoding approach to NCCs, as proposed by Haynes (2009), rests on a conception of NCCs as "necessary for a specific conscious experience" (Haynes, 2009, p. 1), whereas the analysis here is predicated on Chalmers' (2000) definition of NCCs, which we now explain.

2. The Canonical Definition

The idea of Neural Correlates of Consciousness (NCCs) has a surprisingly long history. Fink (2016, 2020) traced first occurrences of the idea back to the 19th century (Gurney, 1881; Marshall, 1901), and reports that it was mentioned in Ward's (1911) entry on 'Psychology' in the Encyclopedia Britannica. The exposition of the concept by Crick and Koch (1990) and Crick (1994) hurled it into the focus of a young science of consciousness. But only with the work of Chalmers (2000) did a canonical definition arise that is now the default understanding in the neuroscience of consciousness (Fink, 2016; Lepauvre & Melloni, 2021). This definition, provided on page 31 of (Chalmers, 2000), is:

Def. 1. An NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C, for the corresponding state of consciousness.[1]

The goal of the following sections is to provide an analysis of how NCCs so defined can be uncovered based on empirical investigations. To provide this analysis, we introduce a light formalism that represents the important non-formal concepts in the above definition. This allows us to proceed with the analysis without committing to specific choices of those concepts.

First, we introduce the symbol **E** to denote the set of states of consciousness that have been chosen in a particular study. Here, 'set' is understood in the mathematical sense.

Working with this set has two advantages for the following analysis. First, it allows us to

[1] A variant of this definition that is often cited in the literature was introduced by Mormann and Koch (2007) and Koch et al. (2016). This variant defines an NCC as "the minimum neuronal mechanisms jointly sufficient for any one specific conscious percept" (Koch et al., 2016, p. 308), where "conscious percept" can either be understood as "a particular phenomenal distinction within an experience", giving what (Koch et al., 2016) call *content-specific NCC*, or "conscious experiences in their entirety", giving what Koch et al. (2016) call *full NCC*. This definition is a *special case* of Chalmers (2000)'s definition cited above, because: (a) both phenomenal distinctions and conscious experiences in their entirety are examples of states of consciousness as understood by Chalmers (2000) (cf. subsections 'Being Conscious' and 'Contents of Consciousness' of the section 'States of Consciousness' in (Chalmers, 2000)), (b) mechanisms that are "jointly sufficient for any one specific conscious percept" constitute a neural system that is "sufficient for any one specific conscious percept", and (c) the explication of the word "any" in Koch et al. (2016) gives the mapping requirement in Chalmers (2000)'s definition. Crick and Koch (1990) do not give an explicit definition of NCCs, though (Mormann & Koch, 2007) and (Koch et al., 2016) cite this paper when they give the above-mentioned definition. Cf. Footnote 7 of (Fink, 2016) for details on the emergence of the canonical definition in the literature.

proceed without presuming one of the various interpretations of the term 'states of consciousness', but leaves this open to the experimenter's choice. The experimenter may choose a notion that describes whether a subject experiences a stimulus consciously, for example, which is often referred to as a question about the content of consciousness, or a notion that corresponds to total experiences of a subject, or a notion that describes global modes of consciousness, as present in the various sleep stages, for example, or a notion that describes whether a subject is conscious at all. All are viable choices, as far as the following analysis is concerned. The second advantage of working with this set is that it allows us to proceed without making an assumption about which states of consciousness specifically are targeted (or accessible) in a study. This, too, varies from case to case. As a result, the methodology developed in what follows is applicable independently of which understanding and which choice of states of consciousness one deems correct when searching for the NCC.

For simplicity of the following analysis, we assume that **E** comprises a state that describes/corresponds to 'none of the other states of consciousness in **E**'. We denote this state as $e_\varnothing$. This state applies if a system has no experiences at all, or if a system has experiences, but these do not fall under any of the other states of consciousness. An example is the 'not seen' state in contrastive analysis, which indicates that a subject did not experience a chosen stimulus consciously, but is non-committal regarding other aspects of the subject's experience (cf. Section 6).

As a result of including the $e_\varnothing$ state, at any point of time, one of the states of **E** applies: either one of the substantive states of consciousness, or, if not, the state $e_\varnothing$. This choice is a convention about the set **E**, which prevents us from having to consider numerous special cases in Section 5 below.

In what follows, we will speak about the 'activation' of a state of consciousness at a particular time to indicate that the state occurred, was realized, applied, or correctly described the experience of the subject at that time. Nothing hinges on this terminology; we introduce it because it sits nicely with the 'activation' terminology that is often applied to neural states. For the purpose of this paper, 'activation' can be understood in any of the above senses.

Second, in order to address the neural systems mentioned in Definition 1, we denote by **Sys** a class of subsystems of the brain that is relevant for a particular study. The typical example of this class in the context of NCCs is a set of regions of interest (ROIs) as specified in a brain atlas. The motivation to consider an abstract class, rather than sets of ROIs alone, is that the methodology proposed below can be applied to a range of different conceptions of subsystems, of which ROIs are but one example. For example, it can be applied to a notion of subsystems as provided by Predictive Processing with Active Inference, as suggested by Hohwy and Seth (2020), or to computationally defined subsystems more generally, as required in the context of Computational Correlates of Consciousness (CCCs) (Cleeremans, 2005; Reggia, Katz, & Davis, 2019), cf. Section 9. For brevity, we will refer to elements S of the class **Sys** simply as 'systems', as it is clear that the systems in questions are subsystems of the brain.

In order to be able to address minimality among neural systems, as required by Definition 1, we assume that the class of subsystems admits of a partial order relation, which we denote as '$\leq$'. For example, when **Sys** consists of ROIs, the partial order is given by inclusion of sets: If $S, S' \in$ **Sys** denote two ROIs, we have $S' \leq S$ iff the ROI S' is a subset of the ROI S.

Finally, in order to address the states of neural systems in Definition 1, we introduce the following notation. For every system S of **Sys** we assume that a suitable notion of states has been chosen, and denote the set of all states of S by $\mathbf{St}(S)$. We call these 'states of S', 'system states', or simply 'states' if the context is clear. Individual states of S are denoted by s, so that we have $s \in \mathbf{St}(S)$.

What is crucial here is that even for a single class of subsystems, various different choices of states are possible. That's because different choices of states correspond to different *levels of analysis* of the systems. Such choices range

from coarse-grained system-level states like 'active'/'inactive' to fine-grained states like individual neuronal firing rates. In assuming that for every system a notion of state is available, we assume that a level of description has been chosen, relative to which the methodology is applied in a particular study. The choice of appropriate level of description in an empirical study will usually reflect the available neuroimaging tools, and also depend on the statistical inference methods that are applied. We will discuss the standard in current NCC research in Section 6 below.

In summary, we have introduced three formal quantities that we will continue to work with in what follows: states of consciousness, **E**, a class of subsystems, **Sys**, and states of subsystems, $\mathbf{St}(S)$.

A quantity that we will not formalize in what follows are the conditions C mentioned in Definition 1. These conditions concern the selection of subjects that are included in empirical studies. Chalmers (2000) discusses a range of options and offers recommendations, for example to restrict to "normal brain functioning, allowing unusual inputs and limited brain stimulation" (Chalmers, 2000, p. 31). The choice of which conditions to include or allow likely makes a huge difference to the result of an empirical investigation of NCCs. The analysis and methodology developed in what follows are applicable to any such choice.

In the next section, we introduce a light formalism to describe the co-activation data obtained in an empirical study, subsequent to which we will go through Definition 1 step-by-step to understand how the NCC defined therein can be measured based on such data.

3. Empirical Data

The analysis carried out in what follows presumes data about the 'co-activation', 'co-occurrence', or 'co-realization' of neural states and states of consciousness. This is a general data type which includes, for example, the type of data that is collected in contrastive analysis studies. Following the same rationale as above, we will introduce a light formal notation to reference such data. The analysis (and methodology) developed in what follows can be applied to any data that adheres to this formalism. Explicitly, this includes, among other options:

1. data about which neural states and states of consciousness were *active* together at some point of time in the experiment,
2. data about which neural states and states of consciousness *occurred* together in the experiment,
3. data about which neural states and states of consciousness were *realized* at some point of time in the experiment,
4. data about which neural states and states of consciousness both *applied* to a subject at some point of time in the experiment,
5. data about which neural states and states of consciousness both *describe* the subject correctly at some point of time in the experiment.

The differences between these concepts hail from differences in conceptions of the concept of 'state' and do not matter for the following analysis. What matters is only that the sufficiency requirement in Definition 1 constrains which data can occur (cf. Section 5.1). For simplicity, we will use the term 'co-activation' as a placeholder for any of the above concepts.

In what follows, we first explain the usual form in which data of the above type is available: as time-series, where each element of the series corresponds to a trial in an experiment. But because the time-ordering of the trials does not matter for Definition 1, an easier and more general way to think about the data is in terms of a data set (or data lake) of pairs of neural states and states of consciousness,[2] which is what the formalism we introduce below represents.

Measures of consciousness are means to infer the state of consciousness of a subject in an experiment. They usually make use of reports (e.g. by pressing of a button) or behavioural indicators (e.g. reaction times) to indicate, perhaps retroactively, which experience a subject likely had in a particular trial in the experiment

[2]This does not speak against studying the relevance of temporal dynamics for NCCs by choosing a temporally extended notion of system states. The point here is only that the order of trials in an experiment does not matter as far as the canonical definition of NCCs is concerned, cf. below.

(Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008; Irvine, 2013). For example, if an experiment is comparing cases where a subject experienced a stimulus consciously with cases where it didn't, as usual in the case of contrastive analysis (Section 6), the measure of consciousness is what takes care of the inference of which was the case in which trial of the experiment.

Choosing an appropriate notion of states of consciousness **E** to represent the target of a measure of consciousness under consideration, for every trial t in an experiment where the measure of consciousness is successfully applied, there is a state of consciousness $e(t)$ that represents the result of the inference procedure. Collecting $e(t)$ of all trials t in a study caters to the intuition of a time series mentioned above. In the case of missing data, for example because the application of the measure of consciousness failed or did not provide a meaningful result in a particular trial, the corresponding element $e(t)$ of the time-series is undefined; there is no entry for this trial.

Information about neural activity is provided by neuroimaging tools. Due to the inherently stochastic nature of both neural activity and neuroimaging tools, statistical inference is required to extract meaningful information about the neural state from the raw data provided by the neuroimaging tools. We will discuss statistics in Section 7 and assume, for now, that the raw data from neuroimaging tools can be preprocessed so as to infer, based on a suitable statistical procedure, which neural activity has generated the raw data that has been measured in an experimental trial, and that this neural activity can be represented by a suitably chosen notion of states on the class of subsystems under consideration.

In contrastive analysis studies, one typically obtains neural information for the whole brain, but applies a statistical procedure on a more local basis. For example, when analysing the data from an fMRI scanner (cf. Section 6), one might obtain neural activity for all voxels, but apply the statistical analysis for each individual voxel separately. In contrast, to optimize for generality, here we work with a formal representation of neural activity that provides a neural state for each (relevant) subsystem of the brain separately. In cases where only a global state is available, and when working with ROIs, the local system states can be obtained by restricting the global state to individual ROIs.

Choosing an appropriate notion of system states $\mathbf{St}(S)$, the neural activity that results from statistical pre-processing of the raw neuroimaging data collected in a trial t can be represented as a state $s(t)$. This caters to the intuition of a time-series of neural states. Because in the formalism applied here, a system state is associated to one system $S \in \mathbf{Sys}$, the neural activity measured in an experiment in fact gives rise to a collection of time series, one such series for every system S whose state has successfully been measured.

Taking into account both the neural and experiential states, the empirical data resulting from experimental trials of an empirical study is a collection $(e(t), s_1(t))$, $(e(t), s_2(t))$, etc., of states that were active during that trial t. Here, $e(t)$ is a state of consciousness inferred by the application of a measure of consciousness in trial t, and the $s_i(t)$ are the system states of the subsystems S_i that were successfully measured in trial t. The states of consciousness, system states, and subsystems are elements of **E**, $\mathbf{St}(S_i)$, and **Sys**, respectively, each of which is chosen according to the study under consideration.

The following analysis, and the resulting methodology, can be applied to time-series data of the above form. However, the ordering of trials in an experiment does not matter for Definition 1. The sufficiency requirement in the definition only constrains which neural states and states of consciousness can be active together, not the order in which they have been measured. Because of this, we can actually disregard the time-ordering information in the time series, and simply discuss the set of all pairs of states that were measured as active together during the experiment (including, possibly, the co-activation of temporally extended neural states with temporally extended states of consciousness if corresponding notions of states have been chosen). We call data of this form 'co-activation data'. It can be obtained from

the time-series by ignoring the time-ordering of the elements of the series.

Def. 2. A data set **D** whose elements are of the form

$$(e, s)\ ,$$

where $e \in \mathbf{E}$ is a state of consciousness, where $s \in \mathbf{St}(S)$ is a system state for some system $S \in \mathbf{Sys}$, and where the conjunction of two states into a pair (e, s) indicates that these states were active together at some point of time during the experiment, is called **co-activation data**.

It is important to note that **E**, **Sys** and $\mathbf{St}(S)$ can be chosen suitably to represent an empirical study under consideration, which is why co-activation data is a rather general type of data that can encompass a range of existing studies. For example, as we will show in Section 6, the data used in contrastive analysis is a form of co-activation data.

Co-activation data is an abstraction from time-series data that has a number of advantages. First, it makes the following analysis much simpler as we can simply reference data (e, s) by using the 'element of' symbol '$\in$' when writing $(e, s) \in \mathbf{D}$. Second, it accommodates the possibility that the statistical analysis of the neuroimaging data provides not just a single state, but a range of states that may have caused the data outputted by the scanner. We can simply add each of the states, together with the state of consciousness that was inferred, to the data set. Third, it is easy to take care of missing data: in cases where either the inference of the system state of a system S or the inference of the state of consciousness went wrong, we simply delete the pair from the data set (while potentially keeping other states that have been successfully inferred). And fourth, most importantly, it allows for differing time scales in neural and experiential data. If a state of consciousness persists much longer than the measured neural state, we can simply add multiple elements to the data set, one for each neural state that occurs, with the same state of consciousness. What matters for the following analysis is only that the states in each element were active together.

Co-activation data is a way to represent the results of measurements in an empirical study. For simplicity, if the context is clear, we will refer to co-activation data **D** simply as 'data set **D**' or 'the data'.

4. Analysing the Logic of Measurement

The goal of Section 5 is to analyse under which circumstances NCCs, as defined in Definition 1, can be discovered empirically in experiments that can measure co-activation data **D**.

Here, it is important to note that Definition 1 refers to the actual world, not data: the scope of the definition are all occurrences of the NCC and NCC states—all 'activations' of these states, in the terminology applied here—not just those that have been measured in a particular study.

In order to distinguish co-activation data that is the result of measurement and statistical inference (cf. Section 3) from the co-activations that actually occur in the real world, we follow the statistical tradition to designate the latter as 'statistical population' or simply 'population'. That is, we use the term 'population' to denote the co-activations of neural and experiential states in the actual world.

The population, so conceived, is what Definition 1 targets, and only if the population states of a system N meet the conditions put forward by Definition 1 is this system an NCC according to this definition.

A logical connection between the definition of NCCs and the data measured in an experiment exists only because the latter says something about the population. In our case specifically, the data **D** says something about which system states and states of consciousness were active together at some point of time during the experiment (cf. Definition 2). This connection is what enables the following analysis. Explicitly:

$$(s, e) \in \mathbf{D} \ \Rightarrow \ \begin{array}{l} s \text{ and } e \text{ were} \\ \text{active together} \\ \text{at some point of time} \end{array} \qquad (4.1)$$

Here, the symbol '$\Rightarrow$' represents the usual logical implication (also called material condition), meaning that for all pairs s and e, if

$(s, e) \in \mathbf{D}$, it follows that s and e were active together at some point of time (in the experiment). Another way of putting this, for those preferring to think in terms of realization, rather than activation, would be: if $(s, e) \in \mathbf{D}$, then at some point of time (in the experiment), s and e must both have been realized. This is a statement about the actual world, viz. the population. Hence we may think of (4.1) as a logical relation between data and the population.

The logical relation between data and the population, (4.1), enables us to analyse the requirements of Definition 1 in light of the available data. In practice, of course, both the system states and states of consciousness that actually occur must be inferred based on noisy observations. This might introduce erroneous observations that violate (4.1). In a statistical context, errors that violate (4.1) are called 'Type I errors', also known as false positives. Type I errors are pairs (s, e) of states that are in $\mathbf{D}$, but which weren't active during the experiment. Thus, in statistical inference, the truth of (4.1) is controlled by the Type I error rate, which is the significance level α of a statistical analysis. Cf. Section 7 for details.

In the first part of Section 5, we will analyse what can be said about NCCs as canonically defined based on co-activation data $\mathbf{D}$ from an experiment if (4.1) holds. The result of this analysis is a modal expression: a statement about which systems *can* satisfy the canonical definition.

The crux of the analysis in the second part of Section 5 is that an empirical search for NCCs can in fact achieve a factual statement—it can find out whether a system *is* the NCC—even if data is partial. This is possible if it can be ascertained, either theoretically or by suitable experimental design, that a system is of *full measure*, defined as follows:

Def. 3. A system $N \in \mathbf{Sys}$ is of **full measure** in the data $\mathbf{D}$ if and only if for all $s \in \mathbf{St}(N)$ and all $e \in \mathbf{E}$ that can be active together, $(s, e) \in \mathbf{D}$.[3]

This definition is violated by Type II errors, also known as false negatives: pairs (s, e) of states that can be active together, but which are not in $\mathbf{D}$. That is, in statistical inference, the truth of Definition 3 is controlled by the Type II error rate β. Statistically speaking, therefore, achieving full measurability is a case of minimizing false negatives. Theoretically speaking, it is a case of ensuring that the conditions of the experiment are such that every combination of states of N and states of consciousness can, if co-realizable, be measured in the experiment. Practically, one can check whether this condition is met at least to some degree: a necessary condition for full measurement of N is that all states $s \in \mathbf{St}(N)$ and $e \in \mathbf{E}$ can be found in the data $\mathbf{D}$. Cf. Section 7 for details.

We note that the emphasis in Definition 3 lies on the limitation to states s of a single system N. While it could be impractical to ask for the converse of (4.1) to be true (which holds for all systems S and their states), Definition 3 only concerns a single system N. We also note that the definition proposes an 'if-then' relation: if $s \in \mathbf{St}(N)$ and $e \in \mathbf{E}$ can be active together, then $(s, e) \in \mathbf{D}$. The 'if and only if' statement in the Definition only concerns the label 'full measure'. This label is applied if and only if the 'if-then' relation holds.

Having discussed the logical relation between data and the statistical population, we can now proceed to analyse the logic of measuring NCCs. This analysis proceeds in two steps. First, in Section 5, we provide a logical analysis of how NCCs can be measured. The result is a theorem that establishes how NCCs can be discovered based on co-activation data. Second, in Section 7, we derive a statistical method that

[3]The term 'full measure' is motivated by the measure-theoretic definition of probabilities, where one often speaks of a statement being true 'up to measure zero' to indicate that the statement might not hold on sets whose measure (viz. probability) is zero. Here, the notion of 'full measure' is meant to indicate that the statement holds for all subsets of a sample space, independently of what their probability of occurrence is. This enables us to carry out a logical analysis in Section 5. Later, in Section 7, we will introduce Type II errors so as to deal with the real-world case in which the statement of Definition 3 only holds up to measure β, where β is the Type II error rate. To keep things as simple as possible, we will avoid any measure-theoretic patois in the rest of the paper, other than the choice of term 'full measure' in this definition.

allows to apply the result of Section 5 in practice in light of noisy observations and statistical errors, with precise control over α and β error rates.

5. Measuring NCCs

In this section, we analyse how NCCs can be identified in empirical studies by use of co-activation data. To this end, we analyse the logic of the canonical definition, provided by Chalmers (2000, p. 31), with respect to co-activation data. To avoid too much repetition, in what follows, we will use the term 'NCC' to refer to the canonical definition, i.e. Definition 1.

Theorem 14, provided at the end of this section, is the result of our analysis. It states how NCCs can be discovered based on empirical co-activation data **D**, making use of three formal quantities that are defined in the following analysis. These quantities are denoted as:

$$\mathbf{D} \twoheadrightarrow \mathbf{B}_N(e) \twoheadrightarrow \mathbf{S}_N(e) \twoheadrightarrow \mathbf{M} \twoheadrightarrow \mathbf{NCC}\,,$$

where the symbol 'A $\twoheadrightarrow$ B' indicates that B can be computed once A is available.

The central formal object in our analysis is the set of all states of a system S that have been measured as co-active with a specific state of consciousness e. We denote this set as $\mathbf{B}_S(e)$, indicating both the system S and the state of consciousness e that the set refers to in the notation. This set can be computed once co-activation data **D** is available. Formally, it is defined as

$$\mathbf{B}_S(e) = \left\{ s \in \mathbf{St}(S) \mid (s, e) \in \mathbf{D} \right\},$$

where we have used notation from mathematics that specifies a set in terms of a condition: the condition is stated after the vertical line '|' and applies to what is stated before the vertical line. In the case at hand, the notation states that $\mathbf{B}_S(e)$ is the set of all those $s \in \mathbf{St}(S)$ for which (s, e) is an element of the data **D**. In what follows, we will often consider the case where the system S in question is denoted as N, in which case the above set is denoted as $\mathbf{B}_N(e)$.

5.1. **Sufficiency.** We start our analysis by the requirement that "a given state of [the NCC] N is sufficient, under conditions C, for the corresponding state of consciousness" in Definition 1.

We have already explained the role of the conditions C in Section 2 (selection of subjects). It remains to consider the requirement that "a given state of N is sufficient for the corresponding state of consciousness", which we call 'sufficiency requirement' in what follows.

Denoting a state of consciousness by e and an arbitrary system in **Sys** by N, our first goal is to find those states of N that can be sufficient for e, given the available data. We denote the set of all such states by $\mathbf{S}_N(e)$:

Def. 4. For a given system $N \in \mathbf{Sys}$ and a given state of consciousness $e \in E$, $\mathbf{S}_N(e)$ is the set of all states $s \in \mathbf{St}(N)$ that can be sufficient for the state of consciousness e, given the data **D**.

We emphasize the use of the word 'can' here. We cannot know whether the states in $\mathbf{S}_N(e)$ really are sufficient as the data may only capture a small window of the actual world, in the sense that it may not comprise all pairs (s, e) that can occur together (viz. all pairs in the population); the data only comprises those pairs that have successfully been measured in the experiment that provided the data. Because of this, the definition only requires that nothing, in the available data, stands against the possibility of sufficiency. Formally: a state $s \in \mathbf{St}(N)$ can be sufficient for the state of consciousness e given the available data **D** if and only if no element of **D** violates the sufficiency requirement of s for e. To ask for more is not logically possible at this stage.

The logic of Definition 4 is as follows. The definition singles out those states in $\mathbf{St}(N)$ which can satisfy the sufficiency requirement with respect to a chosen state of consciousness e given the data. Since $\mathbf{S}_N(e)$ denotes the set of all such states, every state s in $\mathbf{S}_N(e)$ can satisfy the sufficiency requirement with respect to e, and every state $s \in \mathbf{St}(N)$ which is not in $\mathbf{S}_N(e)$ cannot satisfy the sufficiency requirement with respect to e. The definition therefore posits an 'if and only if' relation: a state $s \in \mathbf{St}(N)$ is in $\mathbf{S}_N(e)$ if and only if it can

be sufficient for the state of consciousness e, given the available data $\mathbf{D}$.

Sufficiency is a logical condition. Using the symbol '$\Rightarrow$' to denote implication (also called material condition), sufficiency amounts to the following: for a system state $s \in \mathbf{St}(N)$ and a state of consciousness $e \in \mathbf{E}$, the condition 's is sufficient for e' holds if and only if:

$$s \Rightarrow e \tag{5.1}$$

This is a requirement on how the states in $\mathbf{St}(N)$ and states of consciousness in $\mathbf{E}$ relate. The scope of the requirement are all activations/occurrences/realizations of these states in the actual world. Explicitly, the requirement is governed by the truth table for logical implication, which in the case at hand is

$\mathbf{St}(N)$	$\mathbf{E}$	$s \Rightarrow e$
s	e	T
s	e'	F
s'	e	T
s'	e'	T

where e' denotes any state in $\mathbf{E}$ other than e, s' denotes any state in $\mathbf{St}(N)$ other than s, and T and F denote 'true' and 'false', respectively. The truth table shows that (5.1) is always true if the state of N is not s. Furthermore, it is true if the state of N is s and the state of consciousness is e. The statement is false if the state of N is s, but the state of consciousness is not e.

An important note in this context concerns the case of no experience. The truth table provided above explicates the concept of sufficiency: the statement that a state s of N is sufficient for a state of consciousness e just is a statement of the truth table provided above. One might worry, however, that the sufficiency requirement also says something about the case where a state s occurs, but no conscious experience occurs. This case should, according to the intuitive understanding of the sufficiency requirement, not be possible if s is sufficient for e. In order to cover this case as well, we have assumed that the set of states of consciousness $\mathbf{E}$ also comprises a state of consciousness $e_\varnothing$ which corresponds to the case that there either is no experience at all, or that there is an experience which is not described by any of the other states. Because of this state, the intuition at stake here is in fact part of the second row of the truth table for $e' = e_\varnothing$.

Based on the truth table, we can determine which implications requirement (5.1) has for empirical investigations. As indicated above, we express these implications in terms of the co-activations of system states with states of consciousness, but one could also use the concepts of co-realization or co-occurrence to achieve the same goal. Explicitly, we will say that a system state s and state of consciousness e are active together (or 'co-active', for short) if and only if s and e are active at the same time t. Or, in terms of realization or occurrence: a system state s and state of consciousness e are active together if and only if s and e are both realized/both occur at the same time t. When put in these terms, (5.1) is true if and only if:

1. It is not the case that s is active together with a state $e' \neq e$.
2. If s is active (at a particular time), so is e.

The first condition holds because at any point of time, only one state $e \in \mathbf{E}$ applies. Therefore, Condition 1. is equivalent to the second row of the truth table. Condition 2. is equivalent to the first row of the truth table. Since the third and fourth rows are always true, no other conditions apply: sufficiency of s for e does not have implications for states $s' \neq s$.

Having discussed the meaning of (5.1) in terms of the co-activation of system states and states of consciousness in the actual world (viz. in the population), we can proceed to study the implications for empirical studies and co-activation data $\mathbf{D}$ collected therein. The result is given by the following lemma:

Lemma 5. The set $\mathbf{S}_N(e)$ is the set of all states $s \in \mathbf{St}(N)$ for which

$$s \notin \mathbf{B}_N(e') \tag{5.2}$$

for all $e' \neq e$.

Proof. The lemma claims that a state $s \in \mathbf{St}(N)$ is in $\mathbf{S}_N(e)$ if and only if it satisfies (5.2). We prove the 'if' and 'only if' cases consecutively.

Our first task is to show that if a state $s \in \mathbf{St}(N)$ satisfies (5.2), it is in $\mathbf{S}_N(e)$, meaning that it can be sufficient for the state of consciousness e given the data $\mathbf{D}$. This is the case

if and only if no element of **D** violates the sufficiency requirement of s for e. Above, we have shown that s is sufficient for e if and only if Conditions 1. and 2. hold. Therefore, to prove the 'if' case of the lemma, we need to show that if Condition (5.2) holds, neither Condition 1. nor Condition 2. are violated by elements of the data **D**.

Condition 1. states that it is not the case that s is active together with a state $e' \neq e$. Assume that the condition is violated by elements of the data **D**. This is the case if and only if there is at least one $e' \neq e$ such that s and e' are active together, and such that $(s, e') \in \mathbf{D}$. But $(s, e') \in \mathbf{D}$ if and only if $s \in \mathbf{B}_N(e')$. Therefore, (5.2) does not hold. Thus we have established that if Condition 1. is violated by elements of the data **D**, then (5.2) does not hold. By contraposition (taking the converse of this statement), it follows that if (5.2) holds, then Condition 1. is not violated by elements of the data **D**.

Condition 2. states that if s is active (at a particular time), so is e. Assume, again, that this condition is violated by elements of the data **D**. The condition is violated if and only if s is active at a particular time, but not e. Because one state of **E** applies at any point of time, some $e'' \neq e$ must be active at that point of time (perhaps $e'' = e_\varnothing$). Thus, there must be some $e'' \neq e$ such that s and e'' are active together at that point of time. Thus, Condition 2. is violated by some element of the data **D** if and only if s and some $e'' \neq e$ are active together at that point of time and $(s, e'') \in \mathbf{D}$. But $(s, e'') \in \mathbf{D}$ if and only if $s \in \mathbf{B}_N(e'')$. Therefore, by the same logic as above, we have established that if Condition 2. were violated by elements of the data **D**, then (5.2) does not hold. By contraposition, it again follows that if (5.2) holds, then Condition 2. is not violated by elements of the data **D**.

There are no other implications of (5.1). Therefore, we have shown that if (5.2) holds for a state $s \in \mathbf{St}(N)$, no element of **D** violates the sufficiency requirement of s for e, so that $s \in \mathbf{S}_N(e)$ as claimed. This proves the 'if' case of the lemma.

It remains to prove the 'only if' case of the lemma: if a state $s \in \mathbf{St}(N)$ can be sufficient for the state of consciousness e given the data **D**, then (5.2) holds. We prove this by proving the equivalent converse statement: if (5.2) does not hold for a $s \in \mathbf{St}(N)$, then s cannot be sufficient for the state of consciousness e given the data **D**.

If (5.2) does not hold for a $s \in \mathbf{St}(N)$, there is an $e' \neq e$ such that $s \in \mathbf{B}_N(e')$. This is the case if and only if $(s, e') \in \mathbf{D}$. But because of (4.1), this implies that s and e' must have been active together in the experiment. Because $e' \neq e$, this violates Condition 1., so that s cannot be sufficient for e. Therefore, the 'only if' case of the lemma holds as well. □

Lemma 5 allows us to compute $\mathbf{S}_N(e)$ based on $\mathbf{B}_S(e)$, which we can compute based on **D**.

5.2. **Mapping.** Next, we consider the requirement that "there is a mapping from states of N to states of consciousness, where a given state of N is sufficient for the corresponding state of consciousness" in Definition 1. We will abbreviate this requirement as 'mapping requirement' in what follows.

In terms of the terminology applied here, a mapping from the states of N to states of consciousness is a mapping of the form

$$f : \mathbf{St}(N) \to \mathbf{E} \ . \tag{5.3}$$

A mapping of this form exists if we can map every $s \in \mathbf{St}(N)$ to an $e \in \mathbf{E}$.[4] The requirement that is placed upon this mapping in Definition 1 is the sufficiency requirement studied in Section 5.1 above: that "a given state $[s]$ of N is sufficient for the corresponding state $[f(s)]$ of consciousness".

As in the last section, at this stage of development, we can only investigate whether a mapping that satisfies the sufficiency requirement can exist, not whether such mapping actually does exist in the real world. If elements of the data **D** violate the conditions for the existence of such mapping, it cannot exist. If the

[4]Note that the requirement here is that of a mapping, not that of a partial mapping. A partial mapping would only need to map some of the states of $\mathbf{St}(N)$ to **E**, whereas a mapping needs to map all of the states of $\mathbf{St}(N)$ to **E**: it needs to provide a state $f(s) \in \mathbf{E}$ for every $s \in \mathbf{St}(N)$. This emphasizes again the importance of choosing the right 'level' of description of neural systems when providing a notion of states.

elements of the data $\mathbf{D}$ do not violate such conditions, the mapping can exist, as far as the available data goes. Because a mapping of the form (5.3) is nothing but an association of at most one $e \in \mathbf{E}$ to every $s \in \mathbf{St}(N)$, a mapping where every s is sufficient for the corresponding e can exist if and only if a mapping where every s can be sufficient for the corresponding e does exist. Formally, in more precise terms: A mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, can exist if and only if there is a mapping of the form (5.3), where a given state s of N can be sufficient for the corresponding state of consciousness $f(s)$. The following lemma identifies the condition under which this requirement can be met:

Lemma 6. A mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, can exist if and only if the sets $\mathbf{S}_N(e)$ satisfy

$$\bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e) = \mathbf{St}(N)\,. \tag{5.4}$$

Here, we use the symbol $\bigcup$ to denote the union of sets, so that $\bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e)$ is the union of the sets $\mathbf{S}_N(e)$ for all $e \in \mathbf{E}$. Because $\mathbf{S}_N(e)$ is a subset of $\mathbf{St}(N)$, so is the union. The condition the lemma puts forward in (5.4) is that the union is not just a subset of $\mathbf{St}(N)$, but actually gives the whole set $\mathbf{St}(N)$.

Proof. We prove the 'if' and 'only if' cases consecutively.

To prove the 'only if' case, assume that a mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, can exist. This is the case only if there is a mapping (5.3), where a given state of N can be sufficient for the corresponding state of consciousness $f(s)$. Therefore, for every $e \in \mathbf{E}$ that is in the image of the function, the pre-image $f^{-1}(e)$ contains states $s \in \mathbf{St}(N)$ which can be sufficient for e.[5] Therefore we have $f^{-1}(e) \subseteq \mathbf{S}_N(e)$, where '$\subseteq$' denotes the 'subset or equal' relation.[6] Furthermore, for $e \in \mathbf{E}$ which are not in the image of f, the preimage $f^{-1}(e)$ is the empty set. Thus we have

$$\bigcup_{e\in\mathbf{E}} f^{-1}(e) \subseteq \bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e)\,.$$

For the left-hand-side of this expression, we have $\bigcup_{e\in\mathbf{E}} f^{-1}(e) = f^{-1}(\mathbf{E}) = \mathbf{St}(N)$. Here, the first equality holds because the union of all pre-images of elements of a function is the pre-image of the codomain of the function, and the second equality holds because the pre-image of the codomain of a function is the domain of the function. For the right-hand-side of the expression, since $\mathbf{S}_N(e) \subseteq \mathbf{St}(N)$, we have $\bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e) \subseteq \mathbf{St}(N)$. Combining these two expressions with the above, we thus have

$$\mathbf{St}(N) = \bigcup_{e\in\mathbf{E}} f^{-1}(e) \subseteq \bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e) \subseteq \mathbf{St}(N)\,,$$

which implies that

$$\bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e) = \mathbf{St}(N)\,.$$

This proves the 'only if' case of the lemma.

To prove the 'if' case, assume that (5.4) holds. The condition states that the union of all $\mathbf{S}_N(e)$ covers all of $\mathbf{St}(N)$. Therefore, for every $s \in \mathbf{St}(N)$, we can find at least one e such that $s \in \mathbf{S}_N(e)$. This allows us to define a mapping f of the form (5.3) as follows: for every $s \in \mathbf{St}(N)$, we define $f(s)$ to be any $e \in \mathbf{E}$ for which $s \in \mathbf{S}_N(e)$. This definition gives a mapping of the form (5.3) where a given state s of N can be sufficient for the corresponding state of consciousness $f(s)$. Thus there is a mapping of the form (5.3) where a given state s of N can be sufficient for the corresponding state of consciousness $f(s)$, which means that there can be a mapping of the form (5.3), where a given state s of N is sufficient for the corresponding

[5] For a function $f : X \to Y$, the pre-image of an element y of Y, denoted as $f^{-1}(y)$, is the set of all elements of X which map to y. Explicitly, in terms of the notation for specifying a set by a condition introduced at the beginning of Section 5, $f^{-1}(y) = \{x \in X \mid f(x) = y\}$.

[6] To see that a state s that is sufficient for e is a state that can be sufficient for e, consider the converse statement. A state s can be sufficient for e if and only if there is no element $(s, e') \in \mathbf{D}$ which violates the sufficiency requirement of s for e. Therefore, if s is a state that cannot be sufficient for e, there is an element $(s, e') \in \mathbf{D}$ that violates the sufficiency requirement of s for e. That is the case if and only if $e' \neq e$ and s and e are active together at some point of time. But this implies that s is not sufficient for e.

state of consciousness $f(s)$. This proves the 'if' case of the lemma as well. □

One might wonder whether the mapping f in (5.3) should be required to be surjective (also called 'onto'). While this might be a natural requirement (cf. e.g. (Fink, 2016, Footnote 6)), we will not impose it here, as it is not part of the definition provided by Chalmers (2000), and because, due to the sufficiency requirement in Definition 1, the existence of such mapping is a substantial requirement already even if no surjectivity is demanded. (Plus, surjectivity would imply that one NCC can explain all states of consciousness, which might be too much to ask for.)

We should, however, take care of the special state $e_\varnothing$ that we have introduced in Section 2 to describe the situation where none of the other (substantial) states of consciousness apply, including the case where a subject has no experience at all. As things stand, the mapping f could simply map all of $\mathbf{St}(N)$ to this state, which is not intended by the requirement in Definition 1. Therefore, in addition to the bare sufficiency requirement discussed above, we should require the mapping f to map to at least one state other than $e_\varnothing$. Mathematically speaking, this is the requirement that the image of f (which consists of all $e \in \mathbf{E}$ that the function maps to) should contain at least one state other than $e_\varnothing$. In what follows, we will include this requirement when referring to the 'mapping requirement' and 'NCCs'. The next lemma provides the conditions under which the mapping requirement, thus extended, can be met.

Lemma 7. A mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, and which maps to at least one state $e \neq e_\varnothing$, can exist if and only if the sets $\mathbf{S}_N(e)$ satisfy (5.4) and there is at least one $e \neq e_\varnothing$ such that

$$\mathbf{S}_N(e) \neq \varnothing\,. \tag{5.5}$$

This lemma extends the result of Lemma 6 to ensure that the sufficiency requirement of Definition 1 is met by a mapping f that maps to at least one state other than $e_\varnothing$. The new condition that Lemma 7 puts forward in addition to Lemma 6, Condition (5.5), states that there is at least one $e \neq e_\varnothing$ such that $\mathbf{S}_N(e)$ is not empty. Here, '$\varnothing$' denotes the empty set.

Proof. We prove the 'if' and 'only if' cases consecutively, and refer to the proof of Lemma 6.

To prove the 'if' case, assume that the sets $\mathbf{S}_N(e)$ satisfy (5.4) and that there is at least one $e \neq e_\varnothing$ such that $S_N(e)$ is not empty. Denote this e by e'. The existence of e' allows us to place a further requirement on the function f that we define in the proof of the 'if' case of Lemma 6: we require that for at least one $s \in \mathbf{S}_N(e')$, the function satisfies $f(s) = e'$. This requirement can be met because $\mathbf{S}_N(e')$ is not empty. And it does not interfere with the proof of Lemma 6, because in the proof, we have only required $f(s)$ to be *any* e such that $s \in \mathbf{S}_N(e)$. Choosing e' continues to satisfy this requirement. As a consequence of this additional requirement, e' is in the image of f, so that f maps to at least one state $e \neq e_\varnothing$.

To prove the 'only if' case, assume that a mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, and which maps to at least one state $e \neq e_\varnothing$, can exist. Denote this state again by e'. In the proof of the 'only if' case of Lemma 6, we have already shown that for all e, we have $f^{-1}(e) \subseteq \mathbf{S}_N(e)$, where $f^{-1}(e)$ denotes the pre-image. Since f maps to e', e' is in the image of f, so that the pre-image $f^{-1}(e')$ is not empty. From the above, we have $f^{-1}(e') \subseteq \mathbf{S}_N(e')$, so that if $f^{-1}(e')$ is non-empty, $\mathbf{S}_N(e')$ is non-empty as well. Thus there is least one $e \neq e_\varnothing$ such that $S_N(e)$ is not empty, as claimed. □

The conditions that Lemmas 6 and 7 put forward are conditions on the sets $\mathbf{S}_N(e)$ of a system N. The lemmas show that if the sets $\mathbf{S}_N(e)$ of a system N satisfy (5.4) and (5.5), the system N can satisfy the mapping requirement of Definition 1, and vice versa. For brevity, we will use the abbreviation 'N satisfies (5.4) and (5.5)' to designate the case where the sets $\mathbf{S}_N(e)$ of N satisfy (5.4) and (5.5), and introduce a class $\mathbf{M}$ to denote all systems that do:

Def. 8. Let $\mathbf{M}$ denote all systems in **Sys** that satisfy (5.4) and (5.5).

As in the case of Definition 4, this definition posits an 'if and only if' condition: a system $N \in \mathbf{Sys}$ is in $\mathbf{M}$ if and only if it satisfies (5.4) and (5.5). Or equivalently: any system $N \in \mathbf{M}$ satisfies (5.4) and (5.5), and any system $N \notin \mathbf{M}$ does not satisfy both (5.4) and (5.5).

Making use of the class $\mathbf{M}$, we can reformulate the result of the above lemmas as follows:

Corollary 9. $\mathbf{M}$ is the class of systems $N \in \mathbf{Sys}$ for which a mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, and which maps to at least one state $e \neq e_\varnothing$, can exist.

Proof. Let N be a system in $\mathbf{M}$. By definition, this system satisfies (5.4) and (5.5). Lemma 7 shows that if a system satisfies (5.4) and (5.5), then a mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, and which maps to at least one state $e \neq e_\varnothing$, can exist. If, on the other hand, N is a system which is not an element of $\mathbf{M}$ ($N \notin \mathbf{M}$), then either (5.4) or (5.5) do not hold for N. Therefore, according to Lemma 7, no mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, and which maps to at least one state $e \neq e_\varnothing$, can exist. □

Because $\mathbf{M}$ is defined in terms of (5.4) and (5.5), it can be computed based on $\mathbf{S}_N(e)$.

5.3. **Minimality.** It remains to consider the minimality requirement in Definition 1. As explained in Section 2, minimality is defined with respect to a partial order '$\leq$'. Explicitly, x is a minimal element of a set X if and only if for every $y \in X$, $y \leq x$ implies that $y = x$.[7]

In the case of Definition 1, minimality concerns the partial order '$\leq$' on the set of subsystems $\mathbf{Sys}$. The requirement for a system N to be an NCC is that it is minimal among all those systems that satisfy the mapping requirement discussed above. This is the case, according to the definition of minimality, if and only if, for all systems M that satisfy the mapping requirement, $M \leq N$ implies $M = N$. We will refer to this part of Definition 1 as the 'minimality requirement' in what follows.

The class $\mathbf{M}$ introduced above is the set of systems which *can* satisfy the mapping requirement. Definition 1, on the other hand, identifies an NCC as minimal element of the systems that *do* satisfy the mapping acquirement. Hence our first task is to analyse which systems *can* satisfy the minimality requirement of Definition 1.

Importantly, it is not the case that the class of systems which *can* satisfy the minimality requirement of Definition 1 consists only of those systems which are minimal in the class of systems that *can* satisfy the mapping requirement (viz. the class $\mathbf{M}$). Rather, if nothing is known about the available data except (4.1), every system in the class $\mathbf{M}$ can satisfy the minimality requirement in Definition 1. This is the case because, as shown by Lemma 10 below, additional data makes $\mathbf{M}$ smaller. Intuitively speaking, it could be the case that further measurement produces a pair (s, e) that isn't in $\mathbf{D}$, but in the population, and it can be the case that this pair breaks the sufficiency requirement for s, so that the corresponding system drops out of the class $\mathbf{M}$. Therefore, in principle, as far as the data $\mathbf{D}$ is concerned, it can be the case that additional data constraints $\mathbf{M}$ to any of its subsets, so that any system in $\mathbf{M}$ could, as far as the data $\mathbf{D}$ is concerned, be the minimal system that satisfies the sufficiency requirement of Definition 1. In other words, if nothing is known about the data other than (4.1), the search for NCCs has to stop at $\mathbf{M}$.

Fortunately, as explained in Section 4 and shown below, it can be the case that more is known, which allows a search to go much further.

In the following analysis, we will need to compare different data sets $\mathbf{D}$ and $\mathbf{D}'$. To this end, we amend the existing notation as follows. For the data set denoted by $\mathbf{D}$, all notations are as introduced above. But for a data set denoted

[7] More explicitly, in a preorder $\leq$ (which satisfies the axioms of reflexivity and transitivity), an element x of a set X is a minimal element of X if and only if for every $y \in X$, $y \leq x$ implies $x \leq y$. A partial order is a preorder that satisfies the axiom of antisymmetry in addition to the preorder axioms. Antisymmetry requires that for all x and y, if $x \leq y$ and $y \leq x$, then $x = y$ holds. Thus, for a preorder, an element x is a minimal element of a set X if for every $y \in X$, $y \leq x$ implies $x = y$.

by $\mathbf{D}'$, we add a prime to all existing definitions to indicate that they refer to the data set $\mathbf{D}'$ rather than $\mathbf{D}$. Explicitly, this comprises $\mathbf{B}'_N(e)$ (defined at the beginning of this section), $\mathbf{S}'_N(e)$ (Definition 4), and $\mathbf{M}'$ (Definition 8).

Lemma 10. If $\mathbf{D} \subseteq \mathbf{D}'$, then $\mathbf{M}' \subseteq \mathbf{M}$.

Proof. $\mathbf{D} \subseteq \mathbf{D}'$ means that every $(s, e) \in \mathbf{D}$ is also in $\mathbf{D}'$. Hence, for any $N \in \mathbf{Sys}$, any $s \in \mathbf{St}(N)$ and any $e \in \mathbf{E}$, we have

$$\begin{aligned} s \in \mathbf{B}_N(e) &\Leftrightarrow (s, e) \in \mathbf{D} \\ &\Rightarrow (s, e) \in \mathbf{D}' \Leftrightarrow s \in \mathbf{B}'_N(e)\,, \end{aligned}$$

and therefore

$$s \notin \mathbf{B}'_N(e) \Rightarrow s \notin \mathbf{B}_N(e)\,.$$

Therefore, if a $s \in \mathbf{St}(S)$ satisfies $s \notin \mathbf{B}'_N(e')$ for all $e' \neq e$, it also satisfies $s \notin \mathbf{B}_N(e')$ for all $e' \neq e$. In light of Lemma 5, this establishes that any $s \in \mathbf{S}'_N(e)$ is also in $\mathbf{S}_N(e)$, so that we have

$$\mathbf{S}'_N(e) \subseteq \mathbf{S}_N(e)$$

for all $e \in \mathbf{E}$ and any $N \in \mathbf{Sys}$. This implies, first, that any system N that satisfies

$$\bigcup_{e \in \mathbf{E}} \mathbf{S}'_N(e) = \mathbf{St}(N)$$

also satisfies

$$\bigcup_{e \in \mathbf{E}} \mathbf{S}_N(e) = \mathbf{St}(N)\,,$$

because $\bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e) \supseteq \bigcup_{e\in\mathbf{E}} \mathbf{S}'_N(e) = \mathbf{St}(N)$, and $\bigcup_{e\in\mathbf{E}} \mathbf{S}_N(e) \subseteq \mathbf{St}(N)$. And it implies, second, that any system N that satisfies

$$\mathbf{S}'_N(e) \neq \varnothing$$

also satisfies

$$\mathbf{S}_N(e) \neq \varnothing\,.$$

Therefore, any system N which satisfies (5.4) and (5.5) for $\mathbf{D}'$ also satisfies (5.4) and (5.5) for $\mathbf{D}$, or, put in terms of $\mathbf{M}$ introduced in Definition 8,

$$N \in \mathbf{M}' \Rightarrow N \in \mathbf{M}\,,$$

so that $\mathbf{M}' \subseteq \mathbf{M}$ as claimed. □

Lemma 10 shows that further empirical analysis constrains $\mathbf{M}$ to some subset $\mathbf{M}' \subseteq \mathbf{M}$. If nothing is known about $\mathbf{D}$ other than (4.1), nothing is known about which subset $\mathbf{M}'$ can be found by further empirical analysis so that, as far as $\mathbf{D}$ is concerned, any system in $\mathbf{M}$ could be minimal in the resulting $\mathbf{M}'$, and a fortiori satisfy Definition 1.

Fortunately, as explained in Section 4, it can be the case that more is known about $\mathbf{D}$ and its relation to the world, based on theoretical and/or statistical considerations: it can be the case that one can ascertain, for an individual system N and a suitably chosen maximally acceptable Type II error rate (β), that the system is of full measure in $\mathbf{D}$, Definition 3.

In what follows, we analyse what can be said about the minimality requirement and Definition 1 in light of full measurability of individual systems.

To carry out this analysis, as a first step, we consider the case where every system N is of full measure. This is a theoretical assumption, and we emphasize that this is not an assumption that is needed for the main theorem below. We call data for which this is the case 'complete'.

Def. 11. $\mathbf{D}$ is **complete** iff every $N \in \mathbf{Sys}$ is of full measure in $\mathbf{D}$.

The following propositions will allow us to prove the main theorem (Theorem 14) below.

Proposition 12. If N is minimal in $\mathbf{M}$, and $\mathbf{D}$ is complete, then N is an NCC.

We emphasize that the proposition establishes that N is an NCC, not simply that it can be an NCC. This is possible because of the assumption of complete data in this proposition.

Proof. Consider any $N \in \mathbf{M}$. We first show that any $s \in \mathbf{S}_N(e)$ *is* sufficient for e.

According to Lemma 5, a state $s \in \mathbf{St}(N)$ is in $\mathbf{S}_N(e)$ if and only if $s \notin \mathbf{B}_N(e')$ for all $e' \neq e$. This is the case if and only if $(s, e') \notin \mathbf{D}$. Because $\mathbf{D}$ is complete, N is of full measure in $\mathbf{D}$, which is the case if and only if for all $s \in \mathbf{St}(N)$ and all $e \in \mathbf{E}$ that can be active together, $(s, e) \in \mathbf{D}$ (Definition 3). Formally:

$$\begin{array}{l} s \text{ and } e \text{ can be} \\ \text{active together} \end{array} \;\Rightarrow\; (s, e) \in \mathbf{D}$$

The equivalent converse of this statement is:

$$(s, e) \notin \mathbf{D} \;\Rightarrow\; \begin{array}{l} s \text{ and } e \text{ cannot} \\ \text{be active together} \end{array}$$

Above, we have found that a state $s \in \mathbf{St}(N)$ is in $\mathbf{S}_N(e)$ if and only if $(s, e') \notin \mathbf{D}$ for all

$e' \neq e$. Therefore, making use of the converse statement above, it follows that for any $e' \neq e$, s and e' cannot be active together. Therefore, Condition 1. of Section 5.1 holds true.

Furthermore, because we have assumed that there is a state $e_\varnothing \in \mathbf{E}$ which represents the case that none of the other states apply, if s is active, it must be active with some $e'' \in \mathbf{E}$ (perhaps $e'' = e_\varnothing$). In virtue of the last paragraph, this can only be the case for $e'' = e$. Hence, if s is active at a particular time, so must be e. This is Condition 2. of Section 5.1.

The last two paragraphs show that if a state s is in $\mathbf{St}(N)$, it satisfies the two conditions of sufficiency that we have identified in Section 5.1. This establishes that any $s \in \mathbf{S}_N(e)$ *is* sufficient for e.

Next, we show that N satisfies the mapping requirement. Because $N \in \mathbf{M}$, N satisfies condition (5.4) (Definition 8). This condition implies that every state $s \in \mathbf{St}(N)$ is in some set $\mathbf{S}_N(e)$. Above, we have established that any $s \in \mathbf{S}_N(e)$ *is* sufficient for e. Thus, it follows that for every state $s \in \mathbf{St}(N)$, there is at least one $e \in \mathbf{E}$ such that s is sufficient for e. Defining, $f(s)$ to be that state e for every $s \in \mathbf{St}(N)$ provides a a mapping of the form (5.3) where every state s of N *is* sufficient for the corresponding state of consciousness. Thus, there *is* a mapping from states of N to states of consciousness, where a given state of N is sufficient for the corresponding state of consciousness.

Since $N \in \mathbf{M}$, N also satisfies condition (5.5) (cf. Definition 8). Thus, there is at least one $e' \neq e_\varnothing$ such that $\mathbf{S}_N(e') \neq \varnothing$. From the above, it follows that any $s \in \mathbf{S}_N(e')$ *is* sufficient for e'. Thus, there is at least one $e' \neq e_\varnothing$ for which there is a $s' \in \mathbf{St}(N)$ such that s' is sufficient for e'. Defining $f(s') = e'$ in the construction above ensures that e' is in the image of f, so that f maps to at least one $e \neq e_\varnothing$. Therefore, f also meets the additional requirement we have introduced to deal with the state $e_\varnothing$ (cf. Lemma 7). This establishes that there *is* a mapping from states of N to states of consciousness, where a given state of N is sufficient for the corresponding state of consciousness, and which maps to at least one state $e \neq e_\varnothing$.

Because the above holds true for any $N \in \mathbf{M}$, every $N \in \mathbf{M}$ satisfies the mapping requirement of NCCs. Therefore, if N is minimal in $\mathbf{M}$, it is minimal among the systems that satisfy the mapping requirements. But according to Definition 1, this is the case if and only if N is an NCC. □

Proposition 13. If $\mathbf{D} \subseteq \mathbf{D}'$, $N \in \mathbf{M}$, and N is of full measure in $\mathbf{D}$, then $N \in \mathbf{M}'$.

Proof. We first show that for all $s \in \mathbf{St}(N)$ and $e \in \mathbf{E}$, because of the assumptions of the proposition, we have

$$(s, e) \in \mathbf{D} \Leftrightarrow (s, e) \in \mathbf{D}' \ . \qquad (5.6)$$

Because $\mathbf{D} \subseteq \mathbf{D}'$, we have $(s, e) \in \mathbf{D} \Rightarrow (s, e) \in \mathbf{D}'$. This is the '$\Rightarrow$' direction of (5.6). It remains to show the '$\Leftarrow$' direction: $(s, e) \in \mathbf{D} \Leftarrow (s, e) \in \mathbf{D}'$. The equivalent converse of the last statement is $(s, e) \notin \mathbf{D} \Rightarrow (s, e) \notin \mathbf{D}'$. Because N is of full measure in $\mathbf{D}$, for all $s \in \mathbf{St}(N)$ and $e \in \mathbf{E}$, if $(s, e) \notin \mathbf{D}$, s and e cannot be active together. But because of (4.1), this implies that $(s, e) \notin \mathbf{D}'$, for otherwise, if $(s, e) \in \mathbf{D}'$, s and e were active together, which implies that they can be active together. Thus the '$\Leftarrow$' direction of (5.6) holds as well. But (5.6) implies that

$$\mathbf{B}_N(e) = \mathbf{B}'_N(e)$$

for all $e \in \mathbf{E}$. Because of Lemma 5, this identity implies that

$$\mathbf{S}_N(e) = \mathbf{S}'_N(e)$$

for all $e \in \mathbf{E}$. Therefore, if (5.4) and (5.5) hold for the sets $\mathbf{S}_N(e)$, they also holds for the sets $\mathbf{S}'_N(e)$. Since $N \in \mathbf{M}$, (5.4) and (5.5) hold for the sets $\mathbf{S}_N(e)$. Thus they also hold for the sets $\mathbf{S}'_N(e)$, which implies that $N \in \mathbf{M}'$ as claimed. □

We now return to the case of a single data set $\mathbf{D}$. As before, in the following theorem, we use the term 'NCC' to denote NCCs as in Definition 1, which is the canonical definition in the field first provided by (Chalmers, 2000), with the additional requirement that the mapping maps to at least one of the substantial states of consciousness $e \neq e_\varnothing$.

Theorem 14. If N is minimal in $\mathbf{M}$, and if N is of full measure in $\mathbf{D}$, then N is an NCC.

We remark that the theorem again makes a factual statement in using the word 'is'. Even though **M** denotes the systems that *can* satisfy the mapping requirement in light of the available data (cf. Definition 8), because of the assumption that N is of full measure, the theorem is able to establish that a minimal system in **M** *is* an NCC.

We emphasize that the statement of the theorem is relative to the notion of state that has been applied in an empirical study, and to the system class under consideration, as represented by $\mathbf{St}(S)$ and **Sys**, respectively.

Proof. We denote by $\mathbf{D}'$ the target population of an empirical study, viz. the class of all system states and states of consciousness that can be active together. Using the symbol **St** to denote the union of all system states of all systems in **Sys**, this class is given by

$$\mathbf{D}' = \{\, (s, e) \mid s \in \mathbf{St} \text{ and } e \in \mathbf{E} \text{ can be active together} \,\}.$$

Because, for any $S \in \mathbf{Sys}$, the class $\mathbf{D}'$ contains all $s \in \mathbf{St}(S)$ and $e \in \mathbf{E}$ that can be active together, any $S \in \mathbf{Sys}$ is of full measure in $\mathbf{D}'$ (Definition 3). Therefore, $\mathbf{D}'$ is complete (Definition 11).

Furthermore, because of (4.1), we have $\mathbf{D} \subseteq \mathbf{D}'$. This is the case because according to (4.1), if $(s, e) \in \mathbf{D}$, then s and e were active together. Hence s and e can be active together, so that $(s, e) \in \mathbf{D}'$.

Since $\mathbf{D} \subseteq \mathbf{D}'$, Lemma 10 applies, and establishes that $\mathbf{M}' \subseteq \mathbf{M}$.

Next, we make use of Proposition 13. We have already established that $\mathbf{D} \subseteq \mathbf{D}'$. The theorem furthermore assumes that N is minimal in **M**, and that N is of full measure in **D**. Because every minimal element of a set is an element of the set, the first assumption implies that $N \in \mathbf{M}$. Therefore, all assumptions of Proposition 13 are met, which implies that $N \in \mathbf{M}'$.

Because $\mathbf{M}' \subseteq \mathbf{M}$, because $N \in \mathbf{M}'$, and because N is minimal in **M**, N is also minimal in $\mathbf{M}'$. This is so because, according to the definition of minimality, N is minimal in **M** if and only if for all systems $M \in \mathbf{M}$, $M \leq N$ implies $M = N$ (cf. Footnote 7). Since $\mathbf{M}'$ is a subset of **M** ($\mathbf{M}' \subseteq \mathbf{M}$), this statement includes the statement that for all systems $M \in \mathbf{M}'$, $M \leq N$ implies $M = N$. Because $N \in \mathbf{M}'$, this is the case if and only if N is minimal in $\mathbf{M}'$.

As a final step of the proof, we consider Proposition 12. Since N is minimal in $\mathbf{M}'$, and $\mathbf{D}'$ is complete, the proposition applies to $\mathbf{D}'$, and establishes that N is an NCC. □

Theorem 14 concludes the analysis of how NCCs can be measured based on co-activation data. For later reference, we denote the class of systems that result from the application of the result of this analysis as follows.

Def. 15. We denote by **NCC** all systems in **Sys** which are minimal in **M** and of full measure in **D**.

Theorem 14 establishes that any system that results from the application of the result of the analysis carried out above is an NCC as defined in Definition 1, which is the canonical definition in the field first provided by (Chalmers, 2000), and furthermore satisfies the extended mapping requirement (mapping to at least one of the substantial states of consciousness $e \neq e_\varnothing$) that we have introduced in response to the convention regarding the state $e_\varnothing$ introduced in Section 2. In terms of the class **NCC**, Theorem 14 thus establishes the following:

Corollary 16. If $N \in \mathbf{NCC}$, N is an NCC.

Proof. $N \in \mathbf{NCC}$ if and only if N is minimal in **M** and of full measure in **D**. Therefore, Theorem 14 applies and establishes that N is an NCC. □

The definition of the class **NCC** is the result of the analysis of the measurement of NCCs carried out in this section. The theorems and lemmas of this section show how **NCC** is defined in terms of co-activation data. Therefore, **NCC** can be computed if co-activation data from an empirical study is available.

The results of the logical analysis of measurement of NCCs presented above point at a new way to measure NCCs—that is, to a new methodology to measure NCCs that can, perhaps, be applied in empirical studies to support the search for NCCs. The methodology consists of the measurement of co-activation

data, for example by use of the statistical procedures explained in Section 7, and the subsequent calculation of the class **NCC** based on the definitions provided above. In the following sections, we will analyse this methodology further. To this end, in what follows, we will refer to this methodology as *Co-Activation Analysis* (CoAA).

Explicitly, CoAA consists of the measurement of empirical co-activation data **D** and the computation of the class **NCC** defined above. This computation proceeds as follows. First, one computes the sets $\mathbf{B}_N(e)$ defined at the beginning of this section based on **D**. Second, one assesses Condition (5.2) to compute the sets $\mathbf{S}_N(e)$ based on $\mathbf{B}_N(e)$ (Lemma 5). Third, one assesses Conditions (5.4) and (5.5) to compute **M** based on the sets $\mathbf{S}_N(e)$ (Definition 8). Fourth, one determines the minimal systems in **M** and assesses full measurability in **D**.

We summarize CoAA, and explain how it should be applied in practice, in Section 11. But first, we compare the methodology with contrastive analysis in Section 6, and explain how statistical inference can be applied in Section 7.

6. Contrastive Analysis

In this section, we compare the methodology found above, which we preliminarily call Co-Activation Analysis (CoAA), to contrastive analysis. To this end, we first consider, as an example, a somewhat simplified account of a contrastive fMRI analysis. Based on this example, we then analyse the logic of contrastive analysis abstractly and compare it with the logic of CoAA as found above.

Contrastive analysis presumes a binary distinction in conscious experiences, usually whether a subject experiences a particular stimulus consciously, or not, and aims to measure the difference (contrast, cf. below) in neural states between the two conditions. The former case, where the subject experiences the stimulus, is often called the 'seen' case, whereas the latter case, where the subject does not experience the stimulus consciously, is called the 'unseen' case.

In order to make sure that the contrast that results from the analysis reflects only a difference in conscious experience, and not other systematic differences in the seen vs. unseen cases, for example differences in reporting, working memory, decision making, prior expectations, or particular task demands, sophisticated experimental procedures are necessary, both for the presentation of a stimulus and the inference of whether the stimulus was consciously experienced. This is known as the problem of confounders, cf. (Overgaard, 2004) and (Aru, Bachmann, et al., 2012), which we will discuss in detail in Section 8. In practice, this requirement restricts contrastive analysis to near-threshold conditions, where stimuli (and masks) are chosen so that the resulting experiences of the stimuli are close to the experience/no-experience threshold.

In the case of an fMRI analysis, the neural data that feeds into the contrast analysis is data from an fMRI scanner. In every trial of the experiment, an fMRI scan is produced. We represent the activity of the ith voxel that the fMRI scanner provides after a scan by Y_i. It is taken to be a continuous (real) number.

Together with the fMRI scan, in each trial of the experiment, some behavioural data or a report is collected, so as to determine, by use of a measure of consciousness, whether the subject has experienced the stimulus consciously ('seen cases') or unconsciously ('unseen cases'). The result can be encoded in a contrast variable X, where $X = 0$ corresponds to the unseen cases, and $X = 1$ corresponds to the seen cases. For the purpose of the statistical analysis explained below, X is taken to be a continuous variable as well. The neural activity data from the fMRI scanner, the contrast variable data obtained by use of a measure of consciousness, and other covariates such as age or gender are the empirical data that feed into the contrastive analysis, as we now explain.

In order to determine the difference in neural activation between seen and unseen cases in light of noisy measurements and stochastic brain processes, the empirical data is fed into a statistical model, for example a linear regression

model of the form

$$Y_i = b_{i0} + b_{i1} \cdot X + ...$$

where the dots indicate that further terms are added for covariates such as age or gender. Here, b_{i0} and b_{i1} are parameters of the regression model, also called regression coefficients. Ignoring covariates temporarily, these parameters are interpreted as follows. The parameter b_{i0} is the intercept. In the case of linear regression, it is equal to the average activation of voxel i in the case where $X = 0$, viz. the average activation of voxel i in the unseen cases. The parameter b_{i1} is sometimes called slope. It expresses how much the average activation of voxel i differs between the seen cases ($X = 1$) and the unseen cases ($X = 0$).

Once empirical data is available, the regression can be carried out. There are a range of statistical procedures that can be applied, all of which provide the following information:

1. The analysis provides, first, a descriptive estimation of b_{i0} and b_{i1} (and further parameters in case of covariates). Here 'descriptive' emphasizes that the estimation describes the available data (in the sense of descriptive statistics or exploratory data analysis), but may not hold in the population from which the data is drawn—here, the neural states and states of consciousness of the subject(s).
2. Second, in order to secure the analysis against random fluctuation, a regression analysis includes the result of statistical tests of the estimated parameters (and model in general) against null hypotheses. This is a case of statistical inference that allows to draw conclusion about the population from which the data has been drawn. Examples are t-tests of the model parameters. In the case of contrastive analysis of NCCs, the null hypotheses in the statistical tests of the b_{i1} parameters are $b_{i1} = 0$.

There are many other important steps and quantities that feature in a linear regression analysis (e.g. the evaluation of model fit using coefficients of determination, or the global test of the model against the null model in the form of an F-test), but for the purpose of the following analysis, the above information suffices.

The crucial quantities in a contrastive analysis of NCCs are the p-values of the statistical tests of the parameters b_{i1} against the null hypotheses $b_{i1} = 0$ provided in Step 2.. For every voxel i, one such p-value is available. If the p-value for the ith voxel is smaller than the significance level chosen in the study (e.g. $\alpha = 0.05$), the null hypothesis can be taken to be false, not just in the data, but in the underlying population. In other words, for all voxels with a p-value smaller than the significance level (viz. for significant p-values), the data collected in the experiment affords the conclusion that in the underlying population, $b_{i1} \neq 0$. This means that there is a difference in the average activation of voxel i between the seen and unseen cases *in the population*. We do not know how large the difference really is, or whether there also is a difference for the voxels that did not yields significant p-values, but (provided the formal assumptions of the model hold, and up to the uncertainty reflected in the choice of significance level), it can be taken as certain that there is *a* difference.

Because voxels with significant p-values indicate that there is a systematic difference between seen and unseen cases in the population, assuming that the other requirements of the statistical analysis are satisfied, we can summarize the logic of contrastive analysis in the above example as follows:

- If a brain area/region of interest (ROI) includes voxels with significant p-values of the test of the b_{i1} parameter against $b_{i1} \neq 0$, then the ROI is part of an NCC.

It is needless to say that in practice, there are deviations from this logic. Most notably, some of the significant p-values might be excluded from the above logic for theoretical reasons related to confounders. That is, the significant p-values might be taken to be a result of systematic differences due to reporting, working memory, decision making, prior expectations, or particular task demands in the 'seen' vs. 'unseen' cases, rather than a result of the differences of the conscious experiences in those cases. Or they might be interpreted as artifacts due to the measurement process, especially in the case of lone significant voxels. But once the p-values

are corrected for such theoretical insight, the above logic holds.

In practice, this logic is often applied in a visual inspection of a p-value map. A p-value map shows the p-values for all voxels i which have significant p-values. Making use of a visual inspection has a number of advantages, for example regarding the identification of confounders (cf. above) or the interpretation of artifacts. But once the information in the p-value map is corrected correspondingly, the above logic applies: if a ROI includes significant voxels, it is taken to be part of the NCC.[8]

In what follows, we will analyse the above logic mathematically. To do so, we first show that the data used in contrastive fMRI analysis is a form of co-activation data.

Claim 17. The data used in contrastive fMRI analysis is a form of co-activation data.

To see why Claim 17 holds, we first represent contrastive analysis in terms of the objects **E**, **Sys**, and **St**(S) introduced in Section 2 above, and subsequently consider Definition 2.

First, the binary distinction in conscious experience of a stimulus can be described by two states of consciousness, one state e that represents the seen case, and the state $e_\varnothing$ for the unseen case. Together, these two states provide the set **E** of states of consciousness introduced above, $\mathbf{E} = \{e, e_\varnothing\}$.

Second, since the notion of subsystem that contrastive analysis operates on is that of ROIs, as specified by a suitable brain atlas, we may take **Sys** to be that brain atlas. The partial order '$\leq$' on systems is given by inclusion of sets: If $S, S' \in \mathbf{Sys}$ denote two ROIs, we have $S' \leq S$ iff the ROI S' is a subset of the ROI S.

Third, because the fMRI data provides one activity value for each voxel, the notion of state under consideration is that of activity values for each voxel of a ROI. Explicitly, denoting the voxels of a ROI S by S_V, a state is a mapping

$$s : S_V \to \mathbb{R}\,, \tag{6.1}$$

which associates to every voxel i that belongs to the ROI an activity value $s(i)$. For a ROI S, we denote the set of all such states by **St**(S).

The data that feeds into an fMRI analysis consists of activation values Y_i for each voxel i, together with the value of a contrast variable X which indicates 'seen' vs. 'unseen' cases, as described above. For each trial of the experiment, the activation values for the voxels give states s as in (6.1). The 'seen' cases ($X = 1$) correspond to the state of consciousness e and the 'unseen' cases ($X = 0$) correspond to the state of consciousness $e_\varnothing$. Therefore, each trial of the experiment gives a pair

$$(s, e)$$

consisting of a ROI state s as in (6.1) and a state of consciousness e; these states were observed to be active together at some point of time during the experiment. The set of all (s, e) that were observed in the experiment is data of the co-activation type as defined in Definition 2.

Based on this formal representation of the data, the logic of contrastive analysis can be explicated as follows.

Above, we have seen that if a ROI includes voxels with significant p-values (of the tests of the parameters b_{i1} against the null hypotheses $b_{i1} = 0$), then the ROI is part of the NCC according to contrastive analysis. But a ROI includes voxels with significant p-values of the parameter b_{i1} if and only if there is *some* difference between the average values of the seen and unseen cases in the population. This difference is itself a state of the ROI, of the form (6.1). We denote this state by $\tilde{s}$.

It is the state $\tilde{s}$ that underlies the difference between seen and unseen cases (on average) in the population, according to contrastive analysis. The state is activate in seen cases, $\tilde{s} \in \mathbf{B}_S(e)$, and not activate in unseen cases, $\tilde{s} \notin \mathbf{B}_S(e_\varnothing)$.

Referring to the result of a contrastive analysis as 'Contrast NCC', we can summarize this logic as follows:

[8] The voxels whose p-values are above the significance level cannot be interpreted if standard significance testing is applied. That is because non-significant p-values do not imply that there is no difference in activation in the population; they merely show that nothing can be concluded about the difference in these voxels based on the available data. For this reason, the case where one interprets ROIs with no significant voxels as excluded from the NCC, is not, strictly speaking, a permissible conclusion if the standard suite of linear regression models is applied.

Def. 18. The **Contrast NCC** consists of all those systems $S \in$ **Sys** for which there is a state $\tilde{s} \in \mathbf{St}(S)$ such that

$$\tilde{s} \in \mathbf{B}_S(e)\,, \tag{6.2}$$

but

$$\tilde{s} \notin \mathbf{B}_S(e_\varnothing)\,. \tag{6.3}$$

We denote the class of systems that satisfy these conditions by **ConNCC**.

This definition aligns with the widely shared intuition about contrastive analysis as finding those neural systems that are active in seen cases, and not active in unseen case, relative to a shared background activation of both conditions.

The following theorem shows that the Co-Activation Analysis (CoAA) found above improves upon the results of contrastive analysis, as far as the logic of measurement is concerned. As before, **NCC** is the result of CoAA introduced above (cf. Definition 15 and Corollary 16), and **ConNCC** is the result of a contrastive analysis, defined in the definition above. The theorem shows that every system in **NCC** is also in **ConNCC**. This means that Co-Activation Analysis improves upon the result of contrastive analysis, as far as the logic of measurement is concerned.

Theorem 19. **NCC** $\subseteq$ **ConNCC**.

Proof. Let $N \in$ **NCC**. According to Definition 15, $N \in$ **M**. According to Definition 8, because $N \in$ **M**, N satisfies Conditions (5.4) and (5.5). Because of (5.4), there is at least one state $e' \in$ **E** such that $\mathbf{S}_N(e')$ is non-empty. Because of (5.5), and because $\mathbf{E} = \{e, e_\varnothing\}$, this is the case for the state e which represents the 'seen' cases, so that $\mathbf{S}_N(e)$ is not empty. According to Lemma 5, all states $s \in \mathbf{S}_N(e)$ satisfy

$$s \notin \mathbf{B}_N(e') \tag{6.4}$$

for all $e' \neq e$. Since $\mathbf{S}_N(e)$ is non-empty, it follows that there is at least one state $\tilde{s}$ such that

$$\tilde{s} \notin \mathbf{B}_N(e')\,, \tag{6.5}$$

for all $e' \neq e$. Since $e_\varnothing \neq e$, this implies that

$$\tilde{s} \notin \mathbf{B}_N(e_\varnothing)\,. \tag{6.6}$$

Because $N \in$ **NCC**, according to Definition 15, N is of full measure in **D** (Definition 3). Therefore, for all $s \in \mathbf{St}(N)$ and all $e \in$ **E** that can be active together, $(s, e) \in$ **D**. Due to the state $e_\varnothing$, at any point of time, some $e \in$ **E** must be active (cf. Section 2). Therefore, if $\tilde{s}$ is active, it is active together with some $e'' \in$ **E**. Therefore, $\tilde{s}$ and e'' can be active together. Because of N is of full measure in **D**, this implies that $(\tilde{s}, e'') \in$ **D**. But in light of (6.5), this implies that $e'' = e$. Therefore, we have $(\tilde{s}, e) \in$ **D**, which is the case if and only if

$$\tilde{s} \in \mathbf{B}_N(e)\,. \tag{6.7}$$

Conditions (6.6) and (6.7) are Conditions (6.2) and (6.3) of Definition 18. Therefore, it follows that $N \in$ **ConNCC**. With this, we have shown that any system N which is in **NCC** is also in **ConNCC**. Thus, **NCC** $\subseteq$ **ConNCC**, as claimed.[9] □

7. Statistics

In this section, we discuss how statistics can be used in Co-Activation Analysis (CoAA) to control for noisy observations and stochastic brain processes. We consider two alternative approaches. The first builds on linear regression models as used in contrastive analysis. It yields a theorem that improves upon contrastive analysis as currently applied. This theorem gives a methodology for the empirical search for NCCs which is:

- less constrained than the contrastive approach, because it requires less assumptions regarding the neural states in question.
- statistically more powerful, in the sense that it allows a better control of both Type I and Type II errors.

[9] Strictly speaking, in Definition 18 **D** should be taken to denote pairs of neural states and states of consciousness that can be active together—pairs in the population, that is—as this is what the statistical procedure explained above targets. Adding this requirement to Definition 18 would amount to changing the definition to "The **Contrast NCC** consists of all those systems $S \in$ **Sys** for which in the population $\mathbf{D}'$, there is a state ...". Theorem 19 still applies to this case as well, because if there is a state $\tilde{s}$ in the data that satisfies (6.2) and (6.3) in the data, due to (4.1), and because every $N \in$ **NCC** is of full measure in **D**, the state must also meet those requirements in the population.

- easy to apply in practice.

The second approach makes use of likelihood functions as a bridge between more abstract neural notions and neuroimaging data. It is more general than the first one, albeit also less common and more involved to apply in practice. Crucially, though, it allows to bypass the choice of states that both contrastive analysis and linear-regression CoAA make use of. For reasons of space, and because the details depend on the likelihood function under consideration, the discussion of this second approach is shorter and more high-level than the discussion of the first approach.

There are many other statistical tools that could be used in CoAA as well, which however for reasons of space we do not discuss here. A prime example are Bayesian approaches.

7.1. **Linear Regression-based CoAA.** A crucial question in the search for NCCs is which notion of *state* to consider. The choice of state determines the level or scale at which one is searching for NCCs, and is intimately related to the level or scale of the brain that matters for conscious experience.

Contrastive analysis is based on a linear regression model that amounts to comparing average activity values, where the averages are taken with respect to the 'seen' and 'unseen' cases. In this section, we show that the same type of state can also be subjected to CoActivation Analysis. Doing so leads to a particularly simple and statistically powerful procedure, and in contrast to contrastive analysis, no near-threshold condition or restriction to binary states is required.

To keep the intuition of the following presentation as simple as possible, and to avoid having to introduce too much abstract notation, we consider, as in Section 6, the case of fMRI data. The translation to other experimental schemes should be straightforward.

That is, we assume that there is a set V of voxels which we denote by i. An fMRI scan of the brain provides activity values, which for the ith voxel, we denote as Y_i. As before, we assume these are real numbers $\mathbb{R}$. In conjunction with the fMRI scan, a measure of consciousness is applied to infer information about which state of consciousness e from a set of states of consciousness $\mathbf{E}$ is active. Finally, we assume that covariates such as age or gender are collected too.

We denote the average activity value in voxel i when a state of consciousness $e \in \mathbf{E}$ is active by μ_e^i. That is, μ_e^i is the average of all activity values Y_i which were observed when the state of consciousness inferred by the measure of consciousness was e, adjusted for differences in covariates. If $\mathbf{E}$ is chosen to comprise the 'seen' and 'unseen' cases of contrastive analysis, as indicated by the contrast variable X in Section 6, differences between the μ_e^i are precisely what is investigated in contrastive analysis.

As in the case of contrastive analysis, we take the class **Sys** to comprise a suitable choice of Regions of Interest (ROIs). For a ROI S, a state then consists of the average activity values for each of its voxels. If the state of consciousness relative to which the average is calculated is e, we denote the state of the ROI by s_e. It is a mapping which associates to each voxel i of the ROI the corresponding average activity value, formally:

$$\begin{aligned} s_e : S_V &\to \mathbb{R} \\ i &\mapsto \mu_e^i \,, \end{aligned} \tag{7.1}$$

where S_V denotes the voxels of the ROI S.

Choosing average activity values as the neural states that are considered means setting the state space of every ROI $S \in \mathbf{Sys}$ to consist of the states s_e so constructed.[10] Formally,

$$\mathbf{St}(S) = \{s_e \mid e \in \mathbf{E}\} \,. \tag{7.2}$$

[10] This is a comparably strong assumption because the averages are taken with respect to states of consciousness, which means that the notion of neural state that is presumed is not fully independent from states of consciousness, as one would expect based on the spirit of Definition 1. In the present section, we take this assumption to be warranted by the fact that the statistics of contrastive analysis rely on this assumption too (cf. Section 6), which is why this assumption is deeply baked into the contemporary literature on NCCs. In contrast to contrastive analysis, however, CoAA does not necessitate use of this assumption; Section 7.2 provides one example, among many, of CoAA statistics that do not make use of this assumption.

Prima facie, one might think that one should apply Co-Activation Analysis right to the average activity values, and corresponding states s_e, that are calculated based on the observed data. This, however, would be a mistake, because such procedure would not do justice to the random variations in the activity values brought about by noisy observations and stochastic brain processes. Such random variations make it very likely that the numerical values of the μ^i_e computed in the experiment vary even if the corresponding average activity values in the population are identical. Therefore, just as in the case of contrastive analysis, one needs to make use of statistical inference to obtain information about the underlying statistical population. The information provided by statistical inference is what Co-Activation Analysis makes use of.

In the case we consider here, the crucial inputs for Co-Activation Analysis are differences between neural states in the statistical population. Therefore, one needs to make use of a statistical tool that can infer, based on the measured activity values across trials, whether there are differences in the neural states in the statistical population. As will be shown by Theorem 20 below, Multivariate Analysis of Covariance (MANCOVA) can do the job.

Multivariate Analysis of Covariance is a statistical procedure that tests whether there are differences between multiple dependent variables across various groups when adjusted for covariates. In the case at hand, the dependent variables are the activity values of the voxels i of a ROI S, and, for reasons having to do with Definition 1 (cf. the proof below), the groups comprise pairs e, e' of states of consciousness. If a MANCOVA is significant, we may conclude that there is at least one voxel i of S where $\mu^i_e \neq \mu^i_{e'}$ holds in the population (up to the corresponding Type I error rate α, cf. below). Because of (7.1), this implies that we have $s_e \neq s_{e'}$. If, on the other hand, the MANCOVA is not significant, then up to Type II error rate β, where $(1-\beta)$ is the power of the test, we may conclude that there is no such voxel i. Thus we have $\mu^i_e = \mu^i_{e'}$ for all voxels i of S, which in light of (7.1) implies that we have $s_e = s_{e'}$.[11]

The information so obtained—the information about the difference or identity of the states $s_e \in \mathbf{St}(S)$ *in the population*—decides whether Definition 1 is satisfied, and the Theorems in Section 5 enable us to find out whether this is the case. The difference and identity information constitutes the data that CoAA makes use of in the case at hand.

Because the states s_e are based on the inferred average activity values in the population relative to a state of consciousness e, by definition, a state $s_e \in \mathbf{St}(S)$ is co-active with the state of consciousness e. Therefore, for every $S \in \mathbf{Sys}$, the result of MANCOVA tests can be represented as

$$\mathbf{D} = \{(s_e, e), (s_{e'}, e'), ...\}\,, \qquad (7.3)$$

or $\mathbf{D} = \{(s_e, e) \mid e \in \mathbf{E}\}$ for short. Due to the statistical procedure explained above, the crucial information in this co-activation data is about which states are identical ($s_e = s_{e'}$) and which states are not ($s_e \neq s_{e'}$), according to the statistical analysis under consideration. The explicit numerical values of the μ^i_e delivered by the MANCOVAs are descriptive parameters, but only the inferred differences and identities of states s_e can be generalized to hold in the population.

For later reference, we denote the MANCOVA introduced above by $M^S_{e,e'}$, where S denotes the ROI, and e, e' denote the states of consciousness. Summarized in concise terms, for any $e, e' \in \mathbf{E}$ with $e \neq e'$, we denote by $M^S_{e,e'}$ a MANCOVA whose

- dependent variables are the activity values Y_i of each voxel i of an ROI S,
- whose group variable is $G = \{e, e'\}$,
- and whose covariates are the covariates collected in the experiment.

[11] Because of the control of the Type II error, e.g. by a power analysis, this is an example of the Neyman–Pearson scheme of statistical hypothesis testing, which allows for the interpretation of non-significant results as well.

[12] For notational simplicity, we assume that all MANCOVAs have the same Type I and Type II error rates. A Type I error (false positive) is a case where $M^S_{e,e'}$ is significant, but there is no voxel i of S where $\mu^i_e \neq \mu^i_{e'}$ holds in the population. A Type II error (false negative) is a case where $M^S_{e,e'}$ is not significant, but there is a voxel i of S where $\mu^i_e \neq \mu^i_{e'}$ holds in the population.

We denote the Type I and Type II error rates of these MANCOVAs by α_0 and β_0, respectively.[12]

The following theorem shows how NCCs as defined by Crick and Koch (1990) and Chalmers (2000) (Definition 1) can be discovered with Co-Activation Analysis and the statistical tools introduced above.

Theorem 20. A ROI N is an NCC if and only if it is minimal among all ROIs S that satisfy the following condition: For all $e, e' \in \mathbf{E}$ with $e \neq e'$, the MANCOVA $M^S_{e,e'}$ is significant.

Because MANCOVAs are based on a linear regression model, we refer to the methodology afforded by Theorem 20 as *Linear Regression CoAA*. Bounds on the Type I and Type II error rates for this theorem are given by the following proposition, where α_0 and β_0 denote the Type I and Type II error rates of individual MANCOVAs.

Let J denote the number of pairs in $\mathbf{E}$, and L denote the number of ROIs that are included in (= inside of) the ROI N.[13]

Proposition 21. The probability for a Type I error in Theorem 20 is bounded by

$$\alpha \leq 1 - (1 - \alpha_0)^J \cdot (1 - \beta_0)^L \, .$$

The probability for a Type II error in in Theorem 20 is bounded by

$$\beta \leq \max(\alpha_1, \beta_0) \, ,$$

where $\alpha_1 = 1 - (1 - \alpha_0)^J$.

This proposition gives the overall error rates for the application of Linear Regression CoAA. A Type I error (false positive) of Theorem 20 is a case where the conditions of the theorem are met by N, but N is nevertheless no NCC. A Type II error (false negative) of Theorem 20 occurs if the conditions put forward by the theorem are not met by N, but N is an NCC nevertheless. For notational simplicity, the proposition only states upper bounds for these error rates. A closed formula for the Type I error rate is given in Equation (7.7) below, and more refined bounds for the probability of a Type II error, which depend on how many MANCOVAs fail to be significant, are given in Equation (7.8) ff. below. We will discuss the error rates in Section 11, where we also review how this theorem would be applied in practice.

The remainder of this section is devoted to the proofs of Theorem 20 and Proposition 21. For ease of readability, we separate the logical part of the proofs from the computation of the error rates, and provide each in several steps. We first establish the connection between the condition put forward by the theorem and the set $\mathbf{M}$ in Proposition 22, and subsequently address full measurability in Lemma 24. These statements hold up to the Type I and Type II error rates given above, as established in Lemma 23, Proposition 26, and Proposition 27. The overall proof of the theorem and proposition is a combination of these results, given on page 30.

7.1.1. *Proofs.* The remainder of this section is devoted to the proof of Theorem 20 and Proposition 21, which define Linear Regression CoAA.

Proposition 22. A ROI S is in $\mathbf{M}$ if and only if for all $e, e' \in \mathbf{E}$ with $e \neq e'$, the MANCOVA $M^S_{e,e'}$ is significant.

In determining the error rates for Theorem 20 below, we will often be concerned with Type I and Type II errors of this proposition. A Type I error (false positive) denotes the case where all MANCOVAs $M^S_{e,e'}$ of S are significant, but $S \notin \mathbf{M}$. A Type II error (false negative) denotes the case where at least one of the MANCOVAs is not significant, but we have $S \in \mathbf{M}$ nevertheless.

Proof of Proposition 22. We first prove the 'if' case of the proposition. Therefore, assume that for all $e, e' \in \mathbf{E}$ with $e \neq e'$, the MANCOVA $M^S_{e,e'}$ is significant. This implies, up to Type I errors discussed below, that for each pair e, e' with $e \neq e'$, there is at least one voxel i of S for which we have $\mu^i_e \neq \mu^i_{e'}$. Because of (7.1), this implies that we have $s_e \neq s_{e'}$ for each such pair e, e'.

Making use of (7.3), we can compute the sets $\mathbf{B}_S(e)$ defined in Section 5. They are given

[13] That is, $J = \binom{|\mathbf{E}|}{2}$, where $|\mathbf{E}|$ denotes the cardinality of $\mathbf{E}$ and the brackets indicate the binomial coefficient, and L is the cardinality of the set $\{S \in \mathbf{Sys} | S \leq N, S \neq N\}$, where $N \in \mathbf{Sys}$ is the system designated in Theorem 20.

by

$$\mathbf{B}_S(e) = \{s_e\}$$

for all $e \in \mathbf{E}$.

Lemma 5 states how we can compute the sets $\mathbf{S}_S(e)$ introduced in Definition 4 based on the sets $\mathbf{B}_S(e)$. Since $s_e \neq s_{e'}$ holds for all $e \neq e'$, we have $s_e \notin \mathbf{B}_S(e')$ for all $e' \neq e$. Furthermore, for a given $e \in \mathbf{E}$, all states $s \in \mathbf{St}(S)$ with $s \neq s_e$ are in some $\mathbf{B}_s(e')$ for some $e' \neq e$. Therefore, for a given $e \in \mathbf{E}$, Condition (5.2) holds only for $s = s_e$. We therefore have

$$\mathbf{S}_S(e) = \{s_e\} \tag{7.4}$$

for all $e \in \mathbf{E}$.

In light of the choice of state space in (7.2), this implies that Condition (5.4) of Lemma 6 holds. Since all $\mathbf{S}_S(e)$ are non-empty, Condition (5.5) holds as well, so that in virtue of Definition 8, it follows that $S \in \mathbf{M}$. This establishes the 'if' case of the proposition.

To prove the 'only if' case of the proposition, we assume that the antecedent of the proposition does not hold. That is, we assume that there are at least two states of consciousness $e, e' \in \mathbf{E}$ for which the MANCOVA $M^S_{e,e'}$ is not significant.

Because of the power requirement placed upon the MANCOVA tests, this assumption implies that up to Type II errors discussed below, we have

$$\mu^i_e = \mu^i_{e'}$$

for all voxels i of S. In virtue of (7.1), this implies that

$$s_e = s_{e'} .$$

Since we have $\mathbf{B}_S(e'') = \{s_{e''}\}$ for all $e'' \in \mathbf{E}$, the last equation implies that $\mathbf{B}_S(e) = \mathbf{B}_S(e')$. This, in turn, implies that there is no $s \in \mathbf{St}(S)$ for which (5.2) holds for e or e'. Thus both $\mathbf{S}_S(e)$ and $\mathbf{S}_S(e')$ are empty, and there is no $\mathbf{S}_S(e'')$ that contains either s_e or $s_{e'}$.

Because s_e and $s_{e'}$ are not contained in any $\mathbf{S}_S(e'')$, the left hand side of (5.4) does not contain either of these states. Since the right hand side of (5.4) does include those states, the identity in (5.4) does not hold, so that the system S fails to meet this condition. Therefore, in virtue of Definition 8, it follows that S is not an element of $\mathbf{M}$.

Thus we have shown that if there are at least two states of consciousness $e, e' \in \mathbf{E}$ for which the MANCOVA $M^S_{e,e'}$ is not significant, we have $S \notin \mathbf{M}$. The 'only if' part of the proposition is the contraposition of this statement, so that it holds as well. Thus we have proven the 'only if' part of the proposition as well. □

Next, we provide the error rates for Proposition 22. We have denoted the Type I error of a MANCOVA as α_0, and the Type II error as β_0. Therefore, if a MANCOVA $M^S_{e,e'}$ is significant, we may conclude that $s_e \neq s_{e'}$, but there is a probability of α_0 that this conclusion is wrong, so that in the population $s_e = s_{e'}$ holds. And similarly, because of the Type II error/power control, if a $M^S_{e,e'}$ is not significant, we may conclude that $s_e = s_{e'}$ in the population, but there is a probability of β_0 that this conclusion is wrong, so that we actually have $s_e \neq s_{e'}$ in the population. The following lemma gives the implications of these error probabilities for the proposition above.

Lemma 23. The probability of a Type I error in Proposition 22 is given by

$$\alpha_1 := 1 - (1 - \alpha_0)^J ,$$

where J is the number of pairs in E. The probability of a Type II error in Proposition 22 is given by

$$\beta_S := \beta_0^{K(S)} \cdot (1 - \beta_0)^{J - K(S)} ,$$

where $K(S)$ is the number of MANCOVAs of S that are not significant.

Proof. A Type I error occurs in Proposition 22 if all MANCOVAs $M^S_{e,e'}$ of S are significant, but $S \notin \mathbf{M}$. The proof of Proposition 22 establishes that $S \notin \mathbf{M}$ if and only if for at least one pair e, e', we have $s_e = s_{e'}$. In the case at hand, where all MANCOVAs of S are significant, we have $s_e = s_{e'}$ for at least one pair e, e' if and only if at least one of the MANCOVAs has a false positive (Type I error).

The false positive rate for each MANCOVA is α_0. The number of MANCOVAs required by Proposition 22 is equal to the number of pairs e, e' with $e \neq e'$ in $\mathbf{E}$. Denoting this number by J, the probability of at least one MANCOVA having a false positive is thus given by

$$1 - (1 - \alpha_0)^J .$$

This is the Type I error rate of Proposition 22.

A Type II error (false negative) occurs if at least one of the MANCOVAs is not significant, but we have $S \in \mathbf{M}$ nevertheless, which is the case if and only if we have $s_e \neq s_{e'}$ for all pairs e, e' with $e \neq e'$.

Let us denote the number of MANCOVAs of S that are not significant by $K(S)$. All of these need to have a Type II error (false negative) in order for $s_e \neq s_{e'}$ to hold for all states in the population. Since the Type II error rate for every MANCOVA is β_0, the probability for all of these MANCOVAs to have a Type II error is $\beta_0^{K(S)}$.

All other MANCOVAs are significant and thus need to have true positives for $s_e \neq s_{e'}$ to hold in the population for all e, e'. Each MANCOVA has a true positive with a probability of $1-\beta_0$, and there are $J - K(S)$ such MANCOVAs. The probability that all of these have a true positive is $(1-\beta_0)^{J-K(S)}$.

Combining these two conditions leaves us with a probability of

$$\beta_0^{K(S)} \cdot (1-\beta_0)^{J-K(S)}$$

that $S \in \mathbf{M}$. This is the probability of a Type II error of Proposition 22, which concludes the proof of Lemma 23. □

The next lemma establishes full measurability, defined in Definition 3, up to the error bounds provided by Lemma 23 above.

Lemma 24. If a ROI S satisfies the condition of Theorem 20, then up to the Type I error rate α_1 provided by Lemma 23, S is of full measure in $\mathbf{D}$.

Proof. The definition of states in (7.2) implies that all $s \in \mathbf{St}(S)$ and $e \in \mathbf{E}$ that can be active together are pairs consisting of a state of consciousness $e \in \mathbf{E}$ and the corresponding neural state s_e as defined in (7.1).

S is of full measure in the data $\mathbf{D}$, according to Definition 3, if and only if for all $s \in \mathbf{St}(N)$ and all $e \in \mathbf{E}$ that can be active together, $(s, e) \in \mathbf{D}$. In the case at hand, the data $\mathbf{D}$ is defined in (7.3). The condition of Theorem 20 states that for all $e, e' \in \mathbf{E}$, the MANCOVA $M^S_{e,e'}$ is significant, so that in the data $\mathbf{D}$, we have $s_e \neq s_{e'}$ for all $e, e' \in \mathbf{E}$ with $e \neq e'$.

If for all $e \neq e'$, $s_e \neq s_{e'}$ also holds in the population, then for each pair consisting of an $e \in \mathbf{E}$ and the corresponding $s_e \in \mathbf{St}(S)$ that can be active together, there is an element (s_e, e) in $\mathbf{D}$. Therefore, S is of full measure in $\mathbf{D}$ according to Definition 3.

If, on the other hand, there is a pair e, e' with $e \neq e'$ for which $s_e = s_{e'}$ holds in the population, then the pair s_e and e', as well as the pair $s_{e'}$ and e, can be active together. Because the MANCOVAs are significant, so that we have $s_e \neq s_{e'}$ in the data $\mathbf{D}$, there are no elements (s_e, e') or $(s_{e'}, e)$ in $\mathbf{D}$. This violates full measurability. However, this case only applies if at least one of the MANCOVAs has a Type I error (false positive), the probability of which we have already determined above to be

$$\alpha_1 = 1 - (1-\alpha_0)^J .$$

Up to this Type I error rate, the case before applies. Therefore, up to the Type I error rate α_1 provided by Lemma 23, S is of full measure in $\mathbf{D}$.[14] □

The results above allow us to prove the 'if' part of the theorem.

Proposition 25. A ROI N is an NCC if it is minimal among all ROIs S that satisfy the condition put forward by Theorem 20.

Proof. Proposition 25 is an application of Theorem 14 to statistical inference. Therefore, to prove Proposition 25, we show that the assumptions of Theorem 14 hold, up to the permitted α and β error rates given below.

Proposition 22 establishes that $\mathbf{M}$ is the class of all ROIs which satisfy the condition put forward in Theorem 20. Therefore, the assumptions of Proposition 25 imply that N is minimal in $\mathbf{M}$. This is the first assumption of Theorem 14.

[14] The reason that in the case of Linear Regression CoAA discussed here, full measurability is controlled by the Type I error rate α_1, rather than the Type II error rate β_S, is due to the fact that in the case at hand, the size of $\mathbf{D}$ is fixed, while the number of pairs of states that can be co-active can vary. This number increases if a MANCOVA has a false positive. In more general settings, like the one of Section 7.2, the number of pairs of states that can be co-active is fixed, and the size of the data $\mathbf{D}$ varies. This size decreases if there are false negatives.

Since N satisfies the condition put forward by the theorem, Lemma 24 applies and establishes that N is of full measure in $\mathbf{D}$. This is the second assumption of Theorem 14.

Therefore, N satisfies both assumptions of Theorem 14, so that according to the theorem, N is an NCC. This concludes the proof of Proposition 25. □

Next, we consider the Type I error rate for Theorem 20. This is the overall error rate of the application of Linear Regression CoAA. For notational simplicity, the following proposition only states an upper bound for this error rate. A closed formula for the error rate is given in the proof of the proposition below, cf. Equation (7.7).

Proposition 26. The probability of a Type I error in Theorem 20 is bounded by

$$\alpha \leq 1 - (1 - \alpha_0)^J \cdot (1 - \beta_0)^L \ ,$$

where J is the number of pairs in $\mathbf{E}$, and L is the number of ROIs included in N.

Proof. A Type I error (false positive) of Theorem 20 occurs if the condition of the theorem is met by a system N, but N is nevertheless no NCC. For the purpose of this and the next two proofs, we refer to the condition of Theorem 20 as (C):

(C) N is minimal among all systems S for which all MANCOVAs $M^S_{e,e'}$ are significant.

Here, 'all MANCOVAs $M^S_{e,e'}$' refers to all pairs $e, e' \in \mathbf{E}$ with $e \neq e'$, cf. Theorem 20.

Theorem 14 does not specify conditions that imply that N is not an NCC.[15] Fortunately, though, the case at hand (the case of Linear Regression CoAA) is simple enough to establish Type I error bounds based on first principles.

There are three ways in which N can fail to meet Definition 1: it can fail to satisfy the sufficiency requirement (Section 5.1), the mapping requirement (Section 5.2), or the minimality requirement (Section 5.3).

Regarding the mapping and sufficiency requirements, the crucial question is whether the following condition holds:

(D) $s_e \neq s_{e'}$ in the population for all $e, e' \in \mathbf{E}$ with $e \neq e'$.

As we will show in the next two paragraphs, the sufficiency condition holds if and only if Condition (D) holds, and the mapping condition holds, in the case considered here, if and only if the sufficiency condition holds. Therefore, the probability of N failing to satisfy either of these two conditions even though (C) holds true is the probability that (D) does not hold true even though (C) holds true.

A neural state $s \in \mathbf{St}(S)$ is sufficient for a state of consciousness $e \in \mathbf{E}$ if and only if Condition 1. and 2. on page 10 hold. By construction of the states in Linear Regression CoAA, every state s_e is active together with the corresponding state of consciousness $e \in \mathbf{E}$. If Condition (D) holds, Condition 2. implies that only the state s_e can be sufficient for the state e. And since in this case Condition 1. holds for s_e as well, it follows that every state s_e is sufficient for the corresponding state of consciousness e, and no other state. If, on the other hand, Condition (D) fails, there is at least one pair $e, e' \in \mathbf{E}$ with $e \neq e'$ such that $s_e = s_{e'}$. In this case, s_e is active together with e', and $s_{e'}$ is active together with e. Therefore, there is no state $s \in \mathbf{St}(S)$, where $\mathbf{St}(S)$ is defined in (7.2), which satisfies Condition 1. for e or e'. Thus the sufficiency condition fails. This shows that the sufficiency condition holds if and only if Condition (D) holds.

Regarding the mapping condition, we can make use of the fact that in the case at hand, the number of neural states in $\mathbf{St}(S)$ is equal to the number of states of consciousness in $\mathbf{E}$, by construction. Since the state s_e is sufficient for the state e, every state of consciousness e

[15] It should be possible to derive a further theorem that establishes that N is not an NCC if it fails to meet the conditions of Theorem 14 *with respect to the notion of state specified by the experimenter*, but for reasons of space, we will not do so here. Such theorem should be able to establish the conclusion that N is not an NCC either if $N \notin \mathbf{M}$ (in which case it follows directly from Corollary 5), or, when $N \in \mathbf{M}$, if all systems $\tilde{S} \in \mathbf{M}$ with $\tilde{S} < N$ are of full measure in $\mathbf{D}$. The latter condition guarantees that these $\tilde{S}$ do not fall out of $\mathbf{M}$ as further data is being collected (cf. Lemma 10). It is also important to note, in this context, that even if a system N is found not to be the NCC with respect to the notion of neural states specified by the experimenter, it is still possible for it to be the NCC with respect to a different conception of what the relevant neural states of N are.

has a neural state s_e that is sufficient for it. This implies that a mapping of the form (5.3), where a given state s of N is sufficient for the corresponding state of consciousness $f(s)$, and which maps to at least one state $e \neq e_\varnothing$, exists (cf. Lemma 7). If, on the other hand, the sufficiency requirement fails, so that there is at least one state of consciousness $e \in \mathbf{E}$ for which there is no sufficient neural state, the mapping condition fails as well. Therefore, in the case considered here, the mapping condition holds if and only if the sufficiency condition holds.

Based on these results, we can determine the probability that N fails to satisfy the mapping or sufficiency requirements even though (C) holds true by determining the probability that (D) fails to hold true even though (C) holds true.

Condition (C) includes the condition that all MANCOVAs $M^N_{e,e'}$ of N are significant. Therefore, the probability of (D) being wrong even though (C) is true is the probability that at least one of the MANCOVAs of N has a false positive. Since there are J such MANCOVAs, where J is the number of pairs in $\mathbf{E}$, this probability is

$$\alpha_1 = 1 - (1 - \alpha_0)^J \,, \tag{7.5}$$

as in Lemma 23 above, where α_0 is the Type I error rate for an individual MANCOVA.

But N can still fail to be an NCC even if Condition (D) holds. This is the case if N fails to meet the minimality condition in Definition 1. N fails to meet the minimality condition if there is at least one system $\tilde{S} < N$ which satisfies the sufficiency and mapping conditions as well, where $\tilde{S} < N$ means $\tilde{S} \leq N$ and $\tilde{S} \neq N$.

According to Condition (C), all systems $\tilde{S} < N$ have at least one MANCOVA that is not significant. Thus, for N not to be an NCC, there needs to be at least one such system $\tilde{S}$ that satisfies the mapping and sufficiency requirements nevertheless.

Let us denote the number of MANCOVAs of $\tilde{S}$ that are not significant by $K(\tilde{S})$. All of these need to have a Type II error (false negative) in order for $s_e \neq s_{e'}$ to hold in the population for all pairs e, e'. Since the Type II error rate for every MANCOVA is β_0, the probability for all of these MANCOVAs to have a Type II error is $\beta_0^{K(\tilde{S})}$. All other MANCOVAs need to have true positives. Each MANCOVA has a true positive with a probability of $1 - \beta_0$, and there are $J - K(\tilde{S})$ MANCOVAs that have to have true positives for the above condition to hold, the probability for which is $(1 - \beta_0)^{J - K(\tilde{S})}$. Combining these two conditions leaves us with a probability of

$$\begin{aligned} \beta_{\tilde{S}} &= \beta_0^{K(\tilde{S})} \cdot (1 - \beta_0)^{J - K(\tilde{S})} \\ &= \beta_0^{K(\tilde{S})} - \beta_0^J \,, \end{aligned}$$

as in Lemma 23 above. This is the probability that the system $\tilde{S} < N$ satisfies the mapping and sufficiency conditions even though (C) holds. Here, J denotes the number of pairs in $\mathbf{E}$, as above. For later reference, we note that because, according to (C), it is not the case that all MANCOVAs of $\tilde{S}$ are significant, and because the number of MANCOVAs for any subsystem is bounded by the number of pairs in E, $K(\tilde{S})$ is bounded as

$$1 \leq K(\tilde{S}) \leq J \,.$$

Since $\beta_0 \leq 1$, this implies that $\beta_{\tilde{S}}$ is bounded as

$$\beta_{\tilde{S}} \leq \beta_0 \,. \tag{7.6}$$

Since $\beta_{\tilde{S}}$ is the probability that the system $\tilde{S}$ satisfies the mapping and sufficiency conditions even though (C) holds, the probability that there is at least one system $\tilde{S} < N$ that satisfies the mapping and sufficiency requirements can be expressed as

$$\beta_N := 1 - \prod_{\tilde{S} < N} (1 - \beta_{\tilde{S}}) \,,$$

where the second term denotes the product of $(1 - \beta_{\tilde{S}})$ over all $\tilde{S} < N$.

Given these considerations, we can determine the overall probability that N fails to be an NCC even though (C) holds. To this end, we note that the above cases are independent. There is a probability of α_1 that N does not satisfy the mapping or sufficiency conditions, and a probability of β_N that, if it does satisfy these conditions, it is not minimal. Therefore, the overall probability of N failing to be an NCC if Condition (C) holds is given by

$$\alpha = 1 - (1 - \alpha_1) \cdot (1 - \beta_N) \,.$$

Making use of (7.5) and the formula for β_N, we can express this as

$$\alpha = 1 - (1 - \alpha_0)^J \cdot \prod_{\tilde{S} < N} (1 - \beta_{\tilde{S}}) \,. \tag{7.7}$$

This is the Type I error rate of Theorem 20.

We can derive an upper bound for α by use of Equation (7.6). This equation implies that

$$-(1-\beta_{\tilde{S}}) \leq -(1-\beta_0)\,,$$

so that

$$\alpha \leq 1-(1-\alpha_0)^J \cdot (1-\beta_0)^L\,,$$

where L denotes the number of systems $\tilde{S}$ that are included in N, formally $L := |\{\tilde{S} \in \mathbf{Sys} | \tilde{S} < N\}|$. This proves Proposition 26. □

The following proposition provides an upper bound on the Type II error rate (the probability of a false negative) of Theorem 20.

Proposition 27. The probability of a Type II error in Theorem 20 is bounded by

$$\beta \leq \max(\alpha_1, \beta_0)\,,$$

where α_1 is as in Lemma 23.

Proof. A Type II error (false negative) of Theorem 20 occurs if the conditions put forward by the theorem are not met, but N is an NCC nevertheless. Unlike in the proof of Proposition 26, in this case, we can work with Theorem 14, because it provides conditions which imply that N is an NCC. The theorem's conclusion—that N is an NCC—rests on the following three conditions:

(a) $N \in \mathbf{M}$.
(b) N is of full measure in $\mathbf{D}$.
(c) N satisfies the minimality condition for $\mathbf{M}$.

Here, 'minimality condition' denotes the requirement that no system $\tilde{S}$ with $\tilde{S} < N$ is in $\mathbf{M}$. N is minimal in $\mathbf{M}$ if and only if both (a) and (c) hold. As in the proof of Proposition 26, we denote the conditions put forward by Theorem 20 as (C).

Condition (C) can fail in two ways. First, it can be the case that N is not in the class of systems whose MANCOVAs are all significant. We will denote this case by (C1). Second, even if N is in the class of systems whose MANCOVAs are all significant, it can the case that it is not minimal among systems in that class. We will denote this case by (C2). In what follows, we calculate bounds for the probability that (a) to (c) hold for each of these cases.

In doing so, two facts help to reduce the complexity of the task. First, Condition (b) holds if and only if Condition (a) holds. This is the case because both conditions are true if and only if $s_e \neq s_{e'}$ holds in the population for all $e, e' \in \mathbf{E}$ with $e' \neq e$ (cf. proofs of Proposition 22 and Lemma 24). Therefore, the probability that Conditions (a) and (b) hold even if Condition (C) does not hold can be determined by computing the probability that Condition (a) holds even if Condition (C) does not hold.

Second, in each of the two cases, we can focus on the main driver of the Type II error rate to derive a bound as follows. Let us denote the probability that Conditions (a), (b), and (c) hold by $p_{(\mathrm{a})}$, $p_{(\mathrm{b})}$, and $p_{(\mathrm{c})}$, respectively, and the probabilities that all three conditions hold by $\beta_{(\mathrm{C1})}$ and $\beta_{(\mathrm{C2})}$, where (C1) and (C2) are the cases mentioned above. Since all three conditions have to hold true for N to be an NCC, each of these probabilities is bounded as

$$\begin{aligned}\beta_{(\mathrm{C1})} &\leq p_{(\mathrm{a}|\mathrm{C1})} \cdot p_{(\mathrm{c}|\mathrm{C1})}\\ \beta_{(\mathrm{C2})} &\leq p_{(\mathrm{a}|\mathrm{C2})} \cdot p_{(\mathrm{c}|\mathrm{C2})}\,,\end{aligned}$$

where $p_{(\mathrm{a}|\mathrm{C1})}$ denotes the probability that (a) holds if (C1) is the case, etc.[16] Here, we have already taken the first reduction of complexity mentioned above into account (that (a) holds if and only (b) holds). The product in the formula derives from the fact that (a) and (c) are independent conditions.

The second reduction comes about because probabilities are bounded by 1, so that we have

$$\begin{aligned}\beta_{(\mathrm{C1})} &\leq p_{(\mathrm{a}|\mathrm{C1})}\\ \beta_{(\mathrm{C2})} &\leq p_{(\mathrm{c}|\mathrm{C2})}\,.\end{aligned} \tag{7.8}$$

The probability $p_{(\mathrm{a}|\mathrm{C1})}$ is the probability that Condition (a) holds even though not all of N's MANCOVAs are significant. This is the probability of a Type II error (a false negative) of Proposition 22. It is given by Lemma 23 as

$$\beta_N = \beta_0^{K(N)} \cdot (1-\beta_0)^{J-K(N)}\,,$$

[16] The reason that $\beta_{(\mathrm{C1})}$ is bounded by the product $p_{(\mathrm{a}|\mathrm{C1})}p_{(\mathrm{c}|\mathrm{C1})}$, and not identical to it, is that Conditions (a) to (c) are sufficient for being an NCC, but may not be necessary. Therefore, there may be another set of conditions that is also sufficient for being an NCC, but which has lower Type II error rates.

where $K(N)$ is the number of MANCOVAs of N that are not significant. Thus we have

$$p_{(\mathrm{a}|\mathrm{C1})} = \beta_N \, .$$

In the proof of Proposition 26, we have already shown that due to the bounds on $K(N)$, β_N is bounded as

$$\beta_N \leq \beta_0 \, .$$

Therefore, we have

$$\beta_{(\mathrm{C1})} \leq \beta_0 \, .$$

The probability $p_{(\mathrm{c}|\mathrm{C2})}$ is the probability that (c) holds if N is in the class of systems whose MANCOVAs are all significant, but not minimal among systems in that class. Let us denote the number of systems $\tilde{S} < N$ for which all MANCOVAs $M^{\tilde{S}}_{e,e'}$ are significant by I. We have already denoted the overall number of systems $\tilde{S} < N$ by L, so that the number of systems $\tilde{S} < N$ for which not all MANCOVAs $M^{\tilde{S}}_{e,e'}$ are significant is given by $L - I$. (C2) implies that there are systems $\tilde{S} < N$ whose MANCOVAs are all significant, which is why I is bound as $1 \leq I \leq L$.

The probability that a system $\tilde{S}$ whose MANCOVAs $M^{\tilde{S}}_{e,e'}$ are all significant is not in $\mathbf{M}$ is the probability of a Type I error (false positive) in Proposition 22. It is given by Lemma 23 as

$$\alpha_1 = 1 - (1 - \alpha_0)^J \, .$$

The probability that a system $\tilde{S}$ whose MANCOVAs $M^{\tilde{S}}_{e,e'}$ are not all significant is not in $\mathbf{M}$ is the probability of a true negative in Proposition 22. It is given by

$$1 - \alpha_1 \, ,$$

where α_1 is as above.

N satisfies the minimality condition (c) if no system $\tilde{S}$ with $\tilde{S} < N$ is in $\mathbf{M}$. This is the case if all I systems whose MANCOVAs are all significant have false positives, and all $L - I$ systems whose MANCOVAs are not all significant have true negative cases. Combining the probabilities for these requirements gives an overall probability of

$$p_{(\mathrm{c}|\mathrm{C2})} = \alpha_1^I \cdot (1 - \alpha_1)^{L-I}$$

that N satisfies (c) in case of (C2). Since I is bound as $1 \leq I \leq L$, $p_{(\mathrm{c}|\mathrm{C2})}$ is bound as

$$p_{(\mathrm{c}|\mathrm{C2})} \leq \alpha_1^I \leq \alpha_1 \, .$$

The probability of a Type II error in Theorem 20 depends on whether (C1) or (C2) is the case, and in practice the bound on the Type II error can be computed based on information about which of these cases applies. A general bound that holds for both cases is given simply by the larger of the two elements, viz.

$$\beta \leq \max(\beta_{(\mathrm{C1})}, \beta_{(\mathrm{C2})}) \leq \max(\beta_0, \alpha_1) \, .$$

This proves Proposition 27. □

Having proven these lemmas and propositions, we can now finally turn to the proof of Theorem 20 and Proposition 22.

Proof of Theorem 20 and Proposition 22. We have already proven the 'if' part of Theorem 20 based on Theorem 14 in Proposition 25 above. It remains to prove the 'only if' part.

The 'only if' part of the theorem states that a ROI N is an NCC only if it is minimal among all ROIs S that satisfy the following condition: For all $e, e' \in \mathbf{E}$ with $e \neq e'$, the MANCOVA $M^S_{e,e'}$ is significant. This is equivalent to the statement that if N is an NCC, then N is minimal among all ROIs that satisfy this condition. Or put in terms of Condition (C) introduced above: if N is an NCC, then it satisfies (C).

We can prove this statement based on the result on the Type II error rate we have derived above. To this end, let us first express the statement we have to proof in terms of a truth table, where 'T' denotes true and 'F' denotes false. The truth table for 'if N is an NCC, then it satisfies (C)' is:

N is NCC	(C)	N is NCC $\Rightarrow$ (C)
T	T	T
T	F	F
F	T	T
F	F	T

In Proposition 27, we derived the probability for the following case: N is an NCC, but Condition (C) does not hold. This is the second row of the truth table. The probability that this case applies is β.

The probability that N is an NCC and Condition (C) holds is given by $1 - \beta$. This is the first row of the truth table.

Since the cases in the third and fourth line are always true, by definition of logical implication, it follows that the 'only if' part of the theorem holds up to the Type II error rate stated in Proposition 27 above.[17]

Proposition 25 establishes the 'if' case of the theorem, and Proposition 26 establishes the Type I error rate. Therefore we have proven of Theorem 20 and Proposition 22. □

This concludes the proofs for Linear Regression CoAA.

7.2. **Likelihood-based CoAA.** The statistics developed in the last section apply CoAA to what is, perhaps, the standard and state-of-the-art setting of contemporary contrastive analysis. Doing so brings a number of advantages, for example improved error rates and more flexibility regarding the choice of states of consciousness, cf. Section 10 below.

However, Linear Regression-based CoAA has one big drawback: like contrastive analysis, it analyses a particular notion of *neural* state, namely average activity values. This is due to the linear regression models that are applied in both contrastive analysis (Section 6) and MANCOVA tests (Section 7.1).[18]

In this section, we explain how CoAA can avoid this restriction by working with a different statistical tool: that of a likelihood function that gives the probability of a measurement result (for example, a particular fMRI scanner output) given a brain state under consideration. Such a likelihood function serves as a bridge between a description of the brain in terms of subsystems and states, on the one hand, and the data from a neuroimaging scanner, on the other hand.

Because the likelihood acts as a bridge between brain states and measurement results, a Likelihood-based CoAA is not restricted to the choice of neural states that both contrastive analysis and Linear Regression-based CoAA make. Rather, it can be applied to *any* choice of brain state for which a likelihood function can be given. This might be of advantage for the search for NCCs itself, for example if neural states and neural systems are to be understood in the sense of Predictive Processing, as suggested by Hohwy and Seth (2020), but also for research programmes that aim at finding Computational Correlates of Consciousness (CCCs) or similar proposals, which we will discuss in Section 9.

For reasons of space, and because the concrete mathematics of Likelihood-based CoAA depend on the application, this section constitutes a discussion rather than a presentation of results. The goal is to explain how a Likelihood-based CoAA works. Therefore, unlike Section 7.1 above, this section does not ship ready-to-go results.

7.2.1. *The Task of Statistics in CoAA.* Co-Activation Analysis (CoAA) makes use of data about the co-activation of neural states and states of consciousness. Based on this data, it identifies which system constitutes an NCC (Theorem 14).

In practice, often, the question of which neural state was active at a particular time has to be inferred from measurements made with neuroimaging tools.[19]

The task of statistics in CoAA is to make this inference. That is, statistics take data from neuroimaging scanners across trials as input, and to provide information about which neural states have likely been active in the individual trials as output. Once this information is available, the mathematics of CoAA can do their work.

[17] The crucial part in the 'only if' part of the proof is the bound on the Type II error rate established in Proposition 27. It is always true that a statement like the one here holds up to *some* Type II error.

[18] In contrastive analysis, the linear regression model compares the average activity values of the 'seen' cases with the average activity values of the 'unseen' cases. Linear Regression-based CoAA generalises this so as to take the logic of the definition of NCCs properly into account. However, despite this generalisation, average activity values are still part of the MANCOVA-tests used in Linear Regression-CoAA. As a result, in both contrastive analysis and Linear Regression-based CoAA, the mapping requirement of Definition 1 is already baked into the choice of neural states; in both cases, it holds per assumption. This is why, from a philosophical perspective, Likelihood-based CoAA as described below might be the methodology of choice.

[19] This inference is implicit in both contrastive analysis and Linear Regression-based CoAA because the notion of neural state that both methods presume is determined by the neuroimaging scanner, cf. Footnote 18.

Thus, in a sense, there is a clean division of labour between the mathematics of CoAA, whose task is to make use of the logic of the canonical definition of NCCs to discover whether a system is an NCC, and the statistics, whose task is to create a bridge between the empirical data and the notion of neural state that an experimenter has chosen to investigate. Because of this division of labour, CoAA is compatible with a range of different statistical procedure; its mathematics do not presume a particular choice.

7.2.2. *The Likelihood Function.* A natural choice to build a bridge between the data from neuroimaging scanners and states of consciousness is a likelihood function. If y denotes a measurement result, and s denotes a sate of a neural subsystem S, then a likelihood function is a conditional probability of the form

$$p(y|s)\,,$$

which provides a probability distribution over possible measurement outcomes y, if the neural state is s.

In the case of an fMRI analysis, y would be a vector of activity values y_i of individual voxels that have been observed in a trial, $y = (y_1, y_2, ...)$. The likelihood function would then give the probability of observing this vector if the underlying neural state is s. In many cases, $p(y|s)$ would be a product over individual likelihoods $p(y_i|s)$, which are determined by the choice of states s and general features of the neuroimaging method. If both s and y are real-valued, for example, Gaussian approximations might be applicable in practice.

7.2.3. *Inference of States.* There are several different ways in which a likelihood function can be used to draw inferences about a neural state given measurement outcomes. Depending on details of the likelihood function and measurement outcomes, such inference could be based on:

1. *Maximum Likelihood Methods*, which pick neural states based on which states make the observed data most likely.
2. *Decision Criteria*, which derive a condition that determines whether to accept or reject a neural state as possible cause of the measured data based on α and β error rate considerations.
3. *Bayesian Statistics*, which derive a posterior probability distribution over neural states given observed measurement results based on the likelihood function and a prior probability distribution over neural states.

In practice, instead of working with individual trials, it is likely best to apply these statistical procedures to the set $\mathbf{B}_S(e)$ defined in the beginning of Section 5. This set contains all neural states of the system S that have been co-active with the state of consciousness e, and constitutes the first formal object that CoAA makes use of.

While these explanations surely leave many questions open, we hope that they give at least a first impression of how a Likelihood-based CoAA works. In Section 11.4.2, we discuss how a Likelihood-based CoAA would be applied in practice. Next, we consider one of the deep problems of contrastive analysis, the problem of confounders.

8. Confounders

One of the largest methodological problems in the search for NCCs are confounding factors, also called confounders (Overgaard, 2004; Aru, Bachmann, et al., 2012). In this section, we explore whether Co-Activation Analysis (CoAA) can be of help in resolving the problem of confounders. To this end, we first review what confounders are (Section 8.1) and for which reason they constitute a problem (Section 8.2). Subsequently, we study whether this reason applies to CoAA (Section 8.3), and whether confounders could appear for other reasons (Section 8.4).

8.1. **What are Confounders?** The term 'confounder' is often presented in the terminology of 'factors'. A factor, in an experimental design, is an unobserved variable; and a variable, in this context, indicates a state, mode, configuration, or simply the presence of something, for example of a cognitive function or neural process.

A confounder, in the context of NCC research, then, is a factor that exhibits "any systematic difference between the experimental (conscious) and control (non-conscious) conditions, other than the consciousness itself"

(Overgaard, 2004, p. 220). So any cognitive function or neural process whose state, mode, configuration, or presence exhibits a systematic difference between the conscious and unconscious conditions used in contrastive analysis—the 'seen' and 'unseen' cases explained in Section 6, constitutes a confounder.

Confounders are a deep problem because many experimental protocols needed to measure NCCs carry high risks of introducing confounding factors. For example, the presentation of stimuli alone, or the use of a measure of consciousness to infer whether a subject has experienced a stimulus consciously, may cause processes related to reporting, working memory, or decision making that differ systematically in the seen and unseen cases. Other examples include differences in introspective attitude, attention, or prior expectations (Overgaard, 2004; Melloni, Schwiedrzik, Müller, Rodriguez, & Singer, 2011; Aru, Bachmann, et al., 2012; Aru, Axmacher, et al., 2012; Nani et al., 2019).

Correspondingly, it is no surprise that confounders are generally understood as a problem that needs to be resolved by experimental design. As a consequence, the main focus in research aimed at resolving the issue of confounders, at the present time, is either to devise sophisticated measures of consciousness that allow to side-step systematic differences between the conscious and nonconscious conditions, for example no-report paradigms (Tsuchiya, Wilke, Frässle, & Lamme, 2015; Overgaard & Fazekas, 2016; Block, 2019; Phillips & Morales, 2020; Duman et al., 2022; Hatamimajoumerd, Murty, Pitts, & Cohen, 2022; Dellert et al., 2025), or to make use of experimental strategies to untangle the NCC from confounders, for example by manipulating confounding processes (Aru, Bachmann, et al., 2012) so as to disambiguate which neural activity patterns pertain to confounders as compared to the NCC. Progress in this regard is substantial, but big challenges remain (Lepauvre & Melloni, 2021).

8.2. **Origin of the Problem.** Why do confounders, defined as factors that exhibit a systematic difference between the conscious and unconscious condition, cause a problem?

Confounders are, by definition, factors in an experiment. They are variables that point to cognitive functions or neural processes that appear in an experiment. Therefore, it is natural to think that the experimental protocols that are responsible for the appearance of such functions or processes are the cause of the problem of confounders.

However, the experimental protocols in question do not by themselves imply that processes or functions that exhibit a systematic difference between the conscious and unconscious conditions are a problem. This implication only follows because of the mathematics of contrastive analysis.

Confounders—that is, variables that differ systematically in the 'seen' and 'unseen' cases—constitute a problem only because of a specific mathematical feature that contrastive analysis relies on. For lack of a better term, we may call this feature 'ceteris paribus comparisons'. It is the comparison of two conditions that only differ in the target variable and covariates, but where everything else is assumed to be equal. The statistics of contrastive analysis make use of ceteris paribus comparisons, for example regarding average activity values of an fMRI scan as discussed in Section 6, to identify what is meant to be the NCC.

The problem with ceteris paribus comparisons is that they pick up on *all* significant differences in neural activations that are not accounted for by covariates, independently of what causes these differences. This is the reason why the mathematics of contrastive analysis cannot dissociate between the correlates of differences in consciousness and correlates of differences in other functions or processes, which is why the problem of confounders exist.

When put in causal terms, this shows that the problem of confounders is caused by two separate features of an experiment: experimental protocols that introduce factors that differ systematically between the 'seen' and 'unseen' cases, on the one hand; and ceteris paribus conditions in the mathematics used to analyse the data, on the other hand. It is in this sense that the origin of the problem of confounders comprises both the experimental protocols and the mathematical analysis. Only the combination

of both, confounding factors and mathematics that make use of ceteris paribus comparisons, leads to the problem of confounders as presently understood.

This observation matters for the present purposes because it shows that the problem of confounders may be solved by either change of experimental protocols, or change of the mathematics of the underlying statistical procedure. The latter option does not usually seem to get much attention.

The upshot of this section, then is that in order to resolve the problem of confounders, it might be sufficient to pivot to a analysis method that does not make use of ceteris paribus comparisons. CoAA is such a method, which is why we discuss it next.

8.3. **Confounders and CoAA.** In the last section, we have seen that the origin of the problem of confounders comprises both: experimental protocols that introduce factors that systematically differ between the 'seen' and 'unseen' cases, and statistics which rely on ceteris paribus comparisons that pick up on such differences. Confounders are a problem of experimental and mathematical design, combined.

Therefore, the improvement of experimental protocols is only one of two ways in which one might solve the problem of confounders; a second path to a solution is to improve the mathematics of the statistical analysis under consideration. If a statistical method does not rely on ceteris paribus conditions, one cause of the problem of confounders, as presently understood, ceases to apply.

Co-Activation Analysis (CoAA) does not rely on ceteris paribus comparisons. Rather than finding the NCC by comparing two otherwise equal conditions, it uses the logic inherent in the definition of NCCs to discover the NCC.

As a consequence of this, the mathematical part of the origin of confounders ceases to apply: there is no a priori reason, within CoAA, why factors that exhibit a systematic difference with respect to the 'seen' and 'unseen' cases should confound the result of the analysis. One of the two causes that jointly imply the problem of confounders ceases to apply in the case of CoAA.

This observation is already a good sign that CoAA might be of help in resolving the problem of confounders in experimental practice; because CoAA resolves one of the two causes of confounders in contrastive analysis, one might hope that it can also resolve the problem of confounders entirely.

Whether or not this is the case—whether or not CoAA resolves the problem of confounders in its entirety—depends on whether or not (a) confounders as presently understood in NCC research (factors which exhibit a systematic difference between the 'seen' and 'unseen' cases) can appear for other reasons, and whether or not (b) other types of confounders can appear.

In the next section, we discuss four key differences between CoAA and contrastive analysis that work against these possibilities.

8.4. **Defeating Confounders.** In the last section, we have shown that the reason why confounding factors actually confound an experiment does not apply to CoAA. If CoAA suffers from a problem of confounders, it is for different reasons.

To provide an initial assessment of whether this could be the case, in this section, we study four major differences between contrastive analysis and CoAA, all of which are relevant with respect to the problem of confounders. While the first difference concerns the inner workings of CoAA, the other three differences concern possible applications of CoAA that might resolve remaining issues with confounders, if there are any.

8.4.1. *Logic of NCCs.* The biggest difference between contrastive analysis and CoAA is that CoAA utilizes the logic inherent in the definition of NCCs—the sufficiency, mapping, and minimality requirements discussed in Section 5—, while contrastive analysis doesn't. The work that ceteris paribus comparisons do in contrastive analysis is done by the logic of the definition of NCCs in CoAA.

We have already discussed that one implication of this difference is that the problem of confounders, as presently understood, does not apply to CoAA. The mathematics of CoAA do not make use of what causes the problem of

confounders. Hence, in its known form, the problem doesn't apply.

However, this difference is of much wider importance. Because CoAA makes use of the logic of the definition of NCCs, and because the logic of the definition of NCCs is designed to resolve some of the issues caused by confounders, CoAA has some degree of inherent protection against confounders.

Perhaps the best example thereof concerns what Aru, Bachmann, et al. (2012) have called 'NCC-pr' and 'NCC-co'. 'NCC-pr' denotes neural correlates of prerequisites of conscious perceptions, that is of processes or functions that are causally upstream of conscious experience. 'NCC-co' denotes neural correlates of consequences of conscious perception, that is of processes or functions that are causally downstream of conscious experience, cf. (Bachmann, 2009; Melloni & Singer, 2010; De Graaf, Hsieh, & Sack, 2012). Because such functions or processes exhibit systematic differences between the 'seen' and 'unseen' cases, contrastive analysis is prone to picking up not only the NCC, but also the NCC-pr and the NCC-co. Thus, the NCC-pr and NCC-co are paradigm cases of confounders in contrastive analysis.

But the definition of NCCs provided by Crick and Koch (1990) and Chalmers (2000) (cf. Definition 1) already contains logic to defeat this problem: the sufficiency, mapping, and minimality requirements are made so as to rule out correlates of functions that merely co-vary with consciousness from the NCC. This is most obvious for the minimality requirement, which requires one to pick the smallest ROI among all ROIs that satisfy the other two requirements, hence working against the NCC-pr and NCC-co. But depending on the specific nature of the function and processes in questions, both the sufficiency and mapping requirements might exclude functions or processes that are causally down- or up-stream from conscious experience as well.

Thus the mere fact that CoAA is built on the logic of Definition 1 might already help substantially in resolving the problem of confounders, in addition to the advantage of not having to rely on ceteris paribus comparison statistics.

When viewed from a logical analysis perspective, one could argue that the problem of confounders only exists *because* contrastive analysis does not take the logic of the definition of NCCs into account. The definition of NCCs already protects against some of the problems related to confounders.

8.4.2. *Comparing what?* A second large difference between contrastive analysis and CoAA is that the former is restricted to comparing two states, whereas CoAA is not. CoAA can be applied to any set of states of consciousness, as long as one of them denotes the 'no experience' case (cf. Section 2); it is not limited to two states of consciousness.

This makes a difference for confounders because confounders, as presently understood, are factors that exhibit systematic differences between the 'seen' and 'unseen' states. But there is no reason to think that they necessarily also exhibit notable systematic differences between other states of consciousness. There are good reasons to think that factors that co-vary with the conscious and non-conscious conditions of contrastive analysis do not co-vary significantly with a larger set of states of consciousness.

To provide an illustrating (but oversimplified) example, let us suppose that an experiment is such that there is a cognitive function that is active if and only if a subject experiences a stimulus consciously; perhaps a function related to reporting or working memory. The presence of such function confounds the result of the contrastive analysis because the statistics of contrastive analysis pick up all significant neural differences between the two cases: differences that are due to the difference in conscious experience, but also differences that are due to the difference in said cognitive function.

Consider now, to continue this example, an extension of this experiment to three stimulus conditions. Perhaps a stimulus with a complex mask that results in one case where the stimulus is not experienced at all, and two cases where the stimulus is experienced, in two different ways. Because we have assumed, for the purpose of this example, that the cognitive function is active if and only if a subject experienced a stimulus, it follows that the function is

active in both conditions where the stimulus is experienced. That is to say, the function does not co-vary with the difference between the two conditions; it is invariant with respect to this difference. And because of this, the mathematics of CoAA can remove the neural correlate of the function from the NCC.[20]

8.4.3. *Diversity vs. Uniformity in Conditions.* A third difference concerns the choice of conditions in which subjects can be put across trials. Here, 'condition' comprises the choice of stimulus (as above), but also other factors of the experimental design.

Contrastive analysis, arguably, is premised on having as uniform conditions as possible. The conditions in which subjects are put across trials should be as similar as possible, so as to make sure that the neural differences that the analysis picks up on correspond to differences in conscious experience, and nothing else. Ideally speaking, the conditions are such that the only differences across trials are differences in conscious experience. All additional differences, if systematic, deteriorate the quality of the result of contrastive analysis.

For CoAA, the exact opposite is the case. As we shall explain momentarily, CoAA thrives on diverse, rather than uniform, conditions. Having diverse conditions, including diverse states of consciousness, diverse contexts, or diverse activities that are done during an experiment, seems to enhance the result of the analysis.

This is the case because CoAA is, to a large extent, what has been called an 'exclusionary approach' to NCCs in (Paßler, 2023). CoAA uses information about the co-activation of neural states and states of consciousness to exclude neural states that do not meet the sufficiency or mapping requirements (cf. Section 5.1). As a consequence, the analysis thrives on available information. The more data is available on which neural states are co-active with which states of consciousness, the better. Mathematically speaking, the main reason for this is that Condition (5.2) becomes stronger if data about more neural states is available. Diverse states of consciousness, diverse contexts, diverse activities, of both physical and cognitive form, etc. give rise to more such data, and hence enhance the result of the analysis.

This property of CoAA might be especially helpful for research on NCCs that follows the call for more ecological approaches (Mudrik et al., 2024). Because CoAA is premised on diversity, CoAA might make it easier to measure NCCs in the context of more ecological conditions.

8.4.4. *Different Measures.* A fourth difference, related to the third difference discussed above, is that, logically speaking, it appears as if there is no need for a CoAA based analysis to make use of one and the same measure of consciousness across trials. Logically speaking, it looks as if it is completely fine to combine different measures, or to swap measures in an experiment, or even to combine data from different studies that rely on different measures but the same or similar notions of neural states.

This property of CoAA might help against the problem of confounders because different measures of consciousness may give rise to different confounders. Combining different measures might allow CoAA to disambiguate the NCC against confounding factors of either measure, if these do not overlap.

8.5. **Summary.** In this section, we have argued that CoAA might offer a novel way to resolve the problem of confounders. This is because CoAA differs from contrastive analysis in precisely those matters that cause the problem of confounders, as presently understood. Furthermore, it contains inner logic, inherited from the definition of NCCs by Chalmers (2000) and Crick and Koch (1990), which protects an analysis against confounders that might appear for other reasons to a large extent. And in case a novel notion of confounder still persists despite this protection, CoAA offers several new experimental opportunities to combat confounders further.

It goes without saying that despite these mathematical observations, the application of

[20] The neural trace s_f of the active function f would be excluded by the analysis because s_f fails to satisfy (5.2): it would be in both $\mathbf{B}_S(e)$ and $\mathbf{B}_S(e')$, where e and e' denote the two experiences of the stimulus.

CoAA to real experimental designs is a necessity to exclude confounders with certainty.

9. Computational and Other Correlates of Consciousness

Given all that has been said about Neural Correlates of Consciousness (NCCs) and CoAA so far, one important question remains open: why neurons?

A focus on neurons has, arguably, been crucial in the early phases of the field. With the advent of advanced neuroimaging tools like fMRI scanners, a neuron-centric research agenda for consciousness has, likely, raised the expectation of quick progress that is often crucial in kick-starting new research agendas.

As consciousness science started to become firmly established, however, researchers were quick to point out that the neural level is just one of many levels of interest, and that consciousness might be related to entirely different scales or states. As a consequence, the definition of NCCs has been adapted to other levels of analysis. Important examples include the definition of Computational Correlates of Consciousness by Cleeremans (2005) and the call for an empirical NCC research programme that takes the rich framework provided by the neuroscientific Predictive Processing theory into account (Hohwy & Seth, 2020).

While these proposals have had a substantial impact on empirical research, cf. (Reggia et al., 2019) and (Tal, Wright, Prest, Sandved-Smith, & Sacchet, 2025) for two recent examples, it stands to reason that the full revolution that these proposals may promise has not yet happened. Computational and PP-based methodologies play an important role in theory-building and in models of consciousness, yet there is, so far, no grand research programme that aims at finding theory-independent constraints on how consciousness and computational or Predictive Processing-type states relate.

Whether or not this is due to shortcomings of the available methodologies cannot be assessed within the scope of this paper; next to contrastive analysis, there is the whole suite of decoding-based approaches, after all. But just in case the current methodologies are not entirely satisfying from a computational or PP-based perspective, in this section, we offer a quick review of how CoAA might be of help with these and similar approaches.

9.1. Research Avenues using CoAA.

In our presentation of CoAA in the first part of this paper, we have followed the terminology introduced by Chalmers (2000) when defining NCCs: the terminology of states of consciousness, (neural) systems N, and states of N. In the analysis of the logic of measuring NCCs, which led to CoAA, we made use of a minimal formalization of these notions in terms of the set of states of consciousness $\mathbf{E}$, a class of neural systems $\mathbf{Sys}$ (which carries a partial order '$\leq$' that determines the minimality of systems), and sets $\mathbf{St}(N)$ of system states.

However, as far as the logic and mathematics of CoAA is concerned, nothing hinges on the *neural* interpretation of these formal objects. While neurons provide one valid interpretation of these objects, many other concepts do as well. That is, one can take the quantities $\mathbf{Sys}$ and $\mathbf{St}(N)$ to denote neural subsystems and neural states, but they might equally well denote other conceptions, if these have a similar abstract structure of systems, partial order, and system states.

One noteworthy example for this is the above-mentioned Predictive Processing theory, including its Active Inference form (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017; Parr, Pezzulo, & Friston, 2022; Smith, Friston, & Whyte, 2022), which we will abbreviate as 'PP/ActInf' in what follows. Because PP/ActInf offers a coherent account of action, cognition, and perception, a search for NCCs that utilizes PP/ActInf as the framework of interest might have several large advantages, cf. (Hohwy & Seth, 2020).

With CoAA, such a search can be implemented rather easily. All that is required, from a theoretical perspective, is the specification of the class of subsystems of interest to PP/ActInf, as well as their corresponding states. Based on these choices, a CoAA can be carried out to determine, empirically, whether one of the subsystems is an NCC.

While there is some work involved in defining the required notions, it stands to reason that the empirical work regarding PP/ActInf already makes use of subsystems and states of interest that could be used for a PP-based CoAA. Perhaps all that is needed is a suitable formal definition of these systems. Such a formal definition can, for example, be given in the formal language of category theory, as described in (Tull, Kleiner, & Smithe, 2023), though a simpler mathematical framework, perhaps as in (Buckley, Kim, McGregor, & Seth, 2017), might be suitable as well.

Another noteworthy example are Computational Correlates of Consciousness (CCCs), more broadly understood (Cleeremans, 2005). Since every meaningful conception of computation comes with a set of states and a notion of system, the mathematical quantities that define CoAA—**Sys** and **St**(N)—can easily be obtained.

Depending on the details of the computational states under consideration, it might be possible to run a CCC-based CoAA without a statistical analysis. If no statistical inference of the computational state that is active is required, the definition of CoAA in terms of the sets $\mathbf{B}_N(e)$, $\mathbf{S}_N(e)$, and $\mathbf{M}$ explained in the beginning of Section 5 can be applied "just like that".

If, on the other hand, statistical inference of the computational state is required, it is likely that a Likelihood-based CoAA, as developed in Section 7.2 is the way to go. In this case, one needs a likelihood function that serves as a bridge between the computational brain states, on the one hand, and the neuroimaging data, on the other hand. This function specifies which neuroimaging data one should expect to measure for each of the computational states under consideration.

A third example of interest might be global and dynamical brain states (Mckilliam, 2020; Stevner et al., 2019; Demertzi et al., 2019). Global brain states are states that are distributed over larger brain areas. Dynamical brain states are states that span non-trivial time intervals in neuronal dynamics. As pointed out by (Wiese & Friston, 2021), both constitute a challenge for a ROI-based contrastive analysis.

Due to the mathematical nature of CoAA, neither notion poses a particular problem for CoAA. All that matters is that they can be defined relative to a suitable notion of neural systems, and that this notion of neural systems carries a partial order to make sense of the minimality condition. There are various options of how to do this. For example, for both types of states, one could define neural systems in terms of the support of the states, meaning: the neural systems of interests are those neuronal assemblies over which the states in question are defined, and where they are non-zero, with the partial order given by inclusion. Defining neural systems like this may even make global and dynamical states amenable to a Linear Regression-based CoAA, similar to the one defined in Section 7.1.

It is needless to say that in order to apply CoAA to either of these examples, the details would have to be flashed out. The point of the brief discussion here is only to indicate that CoAA might offer interesting avenues for those who work on NCC-like research programmes with respect to more abstract neural notions, including CCCs, PP/ActInf, and global/dynamical states. Other applications are possible as well. From a mathematical perspective, the limit of CoAA is only the measurability of the corresponding neural states.

10. Methodological Freedom

In the last section, we have already seen that there is considerable freedom in applying Co-Activation Analysis (CoAA), regarding the choice of neural systems and neural states. In this section, we discuss a similar point, but for states of consciousness. Because CoAA does not make any assumptions about which states of consciousness are applied in an experiment, other than that there is a state $e_\varnothing$ that complements the other states of consciousness (cf. Section 2), CoAA can be applied with respect to a wide range of states of consciousness, across various conceptions and/or operationalizations of conscious experience.

This 'methodological freedom' in applications of CoAA derives from the mathematical origin of CoAA. Instead of presuming specific

choices of states of consciousness, neural systems, and neural states, CoAA is built on simple mathematical representations of these notions that preserve an experimenter's freedom to choose what they would like to analyse.

The purpose of this section is to briefly review some examples of states of consciousness that might be of interest to the community and to which, as far as the logic of measurement is concerned, CoAA can be applied. For reasons of space, this review is but a quick random walk through the space of possibilities. Its main function is to illustrate the size of this space, which might be larger than one might otherwise expect.

10.1. **Threshold States.** A first point to briefly mention is that CoAA can be applied to the typical states of consciousness that are considered in contrastive analysis or decoding-based approaches to NCCs. For the present purposes, we will refer to these states as 'threshold states', as they often present a stimulus at the experience/no experience threshold. While working with threshold states introduces a number of challenges, for example regarding the reliability of reports, or possible conflations between unconscious perception and no perception, these states comprise an important pillar of contemporary empirical research programmes.

Because CoAA can be applied to the type of state that contrastive analysis makes use of, mathematically speaking, it can be applied to existing data from contrastive analysis (cf. Section 11.5 for more details).

10.2. **Content States.** An important class of states of consciousness are states that refer to the contents of conscious experience (Hohwy, 2009).

When developing the canonical definition of NCCs, Chalmers (2000) took the content of conscious experience as an important example of states of consciousness, next to a distinction between 'being conscious' and 'not being conscious', background states (cf. below), and states that refer to phenomenal properties in general. The notion of 'state of consciousness' in Chalmers (2000)'s definition (Definition 1) is meant to encapsulate all off these choices, and therefore, so is the formalization of this notion in CoAA.

Content states comprise the threshold states we have discussed in Section 10.1. Therefore, content states may, to a large degree, be amenable to a contrastive analysis. CoAA complements this option by offering an analysis method that can easily be applied to non-threshold content states. This includes states of consciousness of different quantitive and qualitative nature, and non-binary choices of sets of content states.

10.3. **Global States.** A set of states to which contrastive analysis may, arguably, be more difficult to apply are *global states* of consciousness (Mckilliam, 2020), also known as 'modes' of conscious experience, 'overall conscious states' (Hohwy, 2009), or 'background states of consciousness' (Chalmers, 2000). Examples include the states of consciousness that are often referred to in terms of '(alert) wakefulness', 'dreaming', 'dreamless sleep', and similar notions.

Global states pose a problem for contrastive analysis because the conditions in which they typically occur, for example during wakefulness or dreamless sleep, differ across a huge range of factors, not only consciousness, which aggravates the problem of confounders (Section 8). Furthermore, these states cannot be neatly divided into binary 'conscious' vs. 'non-conscious' categories, but appear to exhibit levels or even dimensions (Bayne, Hohwy, & Owen, 2016). This renders these states somewhat incompatible with statistical methodologies aim to identify differences between a conscious and non-conscious condition, such as contrastive analysis.[21]

CoAA, in contrast, can be applied to any notion of states of consciousness that can be represented as a mathematical set $\mathbf{E}$. It is limited only by the measurability of the states

[21] In response to this problem, the contrastive analysis is often applied to a binary choice of global states, either in a 'between-states' paradigm, or in 'within-state' comparisons (Koch et al., 2016). The latter offers a number of noteworthy advantages in terms of confounders (Cecconi et al., 2025), but, fundamentally, the problem still remains (Metzner, Schilling, Traxdorf, Schulze, & Krauss, 2021).

in empirical practice. Since global states can be measured—meaning that it is possible to identify which global state of consciousness a subject experiences during experimental trials– CoAA can be applied to many choices of global states. This includes binary pairs of global states, but also larger sets that comprise various levels of such states. Doing so might allow to side-step the notorious problem of confounders in global states. Mathematically speaking, CoAA might open new research avenues for the global states-based search for NCCs.

One noteworthy example of such research avenues concerns the Neural Correlates of Dreaming (Siclari et al., 2017). Dream states are an important object of investigation for the science of consciousness (Andrillon, 2023; Tononi, Boly, & Cirelli, 2024; Siclari, Patriota, & Olcese, 2025). Since dream states can, arguably, always be represented by a set **E** of states of consciousness, CoAA might help to resolve some of the issues that the contrastive approach faces in this context.

It should be noted, finally, that there are reasons to believe that a strict separation of global states and content states—or of levels and contents of consciousness—might not, ultimately, be tenable (Bachmann & Hudetz, 2014; Hudetz, 2024). From the perspective of CoAA, this does not pose a barrier to experimental investigation; the analysis can be applied to more abstract notions as well.

10.4. **Graded States.** One issue that might appear in the context of both global and content states, but which deserves a separate discussion, concerns graded sates of consciousness.

Conscious experiences, and thus also states of consciousness, may be graded (Jang, Mashour, Hudetz, & Huang, 2024); they might come in degrees (Lee, 2023).

This constitutes an issue for contrastive analysis because the contrastive method in statistics is designed to apply to discrete groups. As a consequence, contrastive analysis, which relies on the contrastive method in statistics, might not be applicable to graded concepts of states of consciousness.

CoAA, at the present stage of its development, relies on a discrete set of states of consciousness as well. As a consequence, it can easily be applied to ordinal notions of states of consciousness—states that have a partial order, that is—but might not be applicable to truly continuous notions of states of consciousness. While the general mathematical formalism introduced in Section 5 might be amenable to continuous states of consciousness, it is unlikely that the simple statistical procedure developed in Section 7.1 can be extended to apply to this case. The likelihood-based procedure explained in Section 7.2 might offer a path forward, but further research is necessary.

10.5. **Micro-Phenomenology.** An exciting recent development in consciousness science is the development of micro-phenomenology (Petitmengin, 2006), a interview method designed to uncover and investigate the fine-grained structure of experience.

Micro-phenomenology has been applied to a wide range of experiences, ranging from lucid dreaming (Demsar & Windt, 2024) or pain (Sparby, Leass, Weger, & Edelhäuser, 2023) to the experience of an intuition (Petitmengin, Remillieux, & Valenzuela-Moguillansky, 2019).

While the investigation of these experiences is an end in its own right, it stands to reason that it can also be very helpful in understanding how conscious experiences relate to the subject matter of the sciences, most notably the brain. In so far as the experiences targeted by micro-phenomenology are examples of minimal phenomenal experience, this is one of the goals of the MPE project (Metzinger, 2020, 2024).

It is unclear whether an NCC research programme is required to meet this goal; perhaps theory-dependent approaches, such as that of computational phenomenology (Ramstead et al., 2022), suffice. But if a theory-independent research programme were helpful—for example a search for NCCs relative to the neural notions provided by Predictive Processing, as explained in Section 9—then a number of problems would need to be resolved.

One problem is that the experiences that micro-phenomenology investigates may not be easily individuated from other experiences

and/or background conditions, which renders these experiences difficult to study in a contrastive approach. A second problem might be that the micro-phenomenological interview method itself introduces a number of confounders; for example, because it makes intricate use of a participant's memory.

It will probably not come as a surprise, this far into this section, that at least as far as these two problems are concerned, the use of CoAA might be an option. CoAA does not make use of ceteris paribus comparisons, which is why it is not necessary, for a CoAA, to individuate or control the experiences in question. And as explained in Section 8, it does not suffer from the problem of confounders in its known form.

Because of this, perhaps, CoAA could be a way to construct a micro-phenomenological NCC research program. While this would surely not be an easy task, when judged from a mathematical perspective, it should be entirely possible.

10.6. **Ecological States.** A final point that might be of interest to those who work on NCCs concerns the important call for more ecological conditions in consciousness science (Mudrik et al., 2024). This call is a reaction to the highly artificial conditions that have, in the past, often been used in the search for NCCs, or in empirical research programmes in consciousness science more generally. These conditions might not appear in ecological contexts, and might therefore lead to the identification of neural mechanisms that are not relevant in practice (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017).

CoAA might be of interest in this context because of two reasons. First, because of its methodological freedom, it might be a helpful tool to search for NCCs in more ecological conditions. As explained above, it does not require an experiment to make use of a contrastive approach (cf. also Section 8), is applicable to a wide range of stimuli or experiences, thrives on diverse rather than uniform conditions across trails (cf. Sections 8.4.3 and 11), and might be compatible with measures of consciousness that would otherwise override the result of a statistical analysis.

But a second advantage might also lie in its utilization of a more intricate logic to identify the NCCs. It could, possibly, be the case that the logic of the analysis helps to avoid some of the wrong identifications that more artificial settings can cause if analysed with the mathematics of a contrastive approach.

For reasons of space, we cannot explore the latter point in detail here. If CoAA is of empirical use at all, this last point will have to be explored in a separate context.

This concludes a brief survey of some of the states of consciousness that CoAA could be applied to. Next, we review how CoAA would be applied in practice.

11. Application & Existing Data

In the previous sections, we have studied the mathematics of Co-Activation Analysis (CoAA) (Section 5), compared CoAA to contrastive analysis (Section 6), developed a first statistical procedure that can be used together with CoAA (Section 7), and discussed some of CoAA's properties that might be helpful in experiments (Sections 8-10). Within these discussions, the details of how one would go about applying CoAA in practice might easily be lost, which is why we review these details here.

It is needless to say that the conception of new experiments is a demanding task, where mathematical and statistical considerations only play a small role. This section, therefore, is in no way meant to provide a systematic guide for the construction of experiments; rather, it is meant to summarize how CoAA can be applied, or may make a difference, to the major dimensions of experimental work.

11.1. **Choice of Analysandum.** A first dimension of an application of CoAA in an experiment concerns the choice of analysandum—the choice of what is to be analysed in an experiment. For CoAA, this is the question of which neural and phenomenal concepts are to be used in the experiment in order to search for the NCC. This includes the following choices:

1. Which **neural systems** should be included in the search for NCCs. A straightforward example of this choice would be ROIs as defined by a suitable brain atlas, but more

abstract choices are possible as well. One example of a more abstract choice would be the neural subsystems in a Predictive Processing framework, as suggested in (Hohwy & Seth, 2020).

2. Which **notion of neural states** should be considered in the search for NCCs. Different choices of states correspond to different levels or scales of analysis, and may depend on the available neuroimaging tools. A straightforward choice in an fMRI analysis, for example, are average activity values per voxel (cf. Section 7.1). Other examples are global or dynamical neural states, cf. Section 9.
3. Which **states of consciousness** are to be used in an experiment. CoAA can be applied to the states that are used in contrastive analysis (e.g. stimuli at the experience/no-experience threshold), but it can also be applied to more general choices, e.g. global states of consciousness, or a larger number of content states. It is also applicable to states that are derived from more ecological conditions or novel methodological approaches (cf. Section 10). The choice of states of consciousness does not need to be binary, and near-threshold conditions are not required.

In practice, these choices might already be determined, to some degree, by the experimental protocols that are to be used: the question of which neural states and neural systems are investigated may be determined by the available neuroimaging tools; and the question of which states of consciousness can be used may be determined by which measures of consciousness can reasonably be applied in an experiment. One advantage of CoAA is that it does not add further requirements to the range of options that the available experimental protocols offer.

If CoAA is to be applied to an experiment that has previously been used in contrastive analysis studies, then no choice of neural system and neural states is required; CoAA can be run with the exact same notions of neural systems and states that contrastive analysis makes use of. This gives rise to a particularly simple statistical procedure, which we will review in Section 11.4.1. The experimenter is still free to choose different states of consciousness, though, including stimuli of different qualitative and quantitative natures.

11.2. **Experimental Design.** A second dimension of the application of CoAA in an experiment concerns the design of the experiment, both in theory and in practice.

The design of an experiment is, of course, not a formal process. It is where much of the creativity and ingenuity of experimenters plays out. What CoAA adds to this process is a range of new empirical research avenues.

On the one hand, these research avenues concern the neural and phenomenal concepts discussed in the last section. CoAA can be applied to more general notions of neural systems, neural states, and states of consciousness than is the case for contrastive analysis.

As important, however, are implications for more subtle aspects of experimental design. One big difference between CoAA and contrastive analysis, for example, concerns the conditions that each analysis presumes. While contrastive analysis is premised on uniform conditions across trials, CoAA thrives on diverse conditions across trials (Section 8.4.3). This might offer new research avenues regarding the contexts, activities, or tasks in experimental design.

11.3. **Data Acquisition.** A third dimension in an application of CoAA concerns the data acquisition.

CoAA relies on co-activation data (Definition 2). This is data about which neural states s are active together with which states of consciousness e. Here, instead of the *activation* of states of consciousness, one could also refer to the *realization*, *occurrence*, or *application* of these states.

Co-activation data is a way to represent the results of measurements in an experiment. It is a general data type, designed to comprise the sort of data that is obtained in contrastive analysis, where in each trial of the experiment, both neural and experiential data is collected.

For CoAA, no time-series information is required. All co-activations of a neural state s

and a state of consciousness e that are observed in an experiment can be added to a data lake $\mathbf{D}$, whose elements are pairs (s, e) of such co-activations. This data lake is what the mathematics of CoAA make use of to identify an NCC.

Working with the data lake $\mathbf{D}$ has a number of advantages, summarized at the end of Section 3. For example, it makes it easy to take differing time scales of experiential and neural states into account.

The only subtlety related to $\mathbf{D}$ is that CoAA requires one to keep track of which neural system a neural state s belongs to, so that the sufficiency requirement in the definition of NCCs can be checked. If a neuroimaging tool outputs a global state, as for example the case for an fMRI scanner, one can simply add multiple co-activations to $\mathbf{D}$ per trial, one for every neural system under consideration.

11.4. **Statistics & Analysis.** This brings us, finally, to the most important dimension of an application of CoAA, the data analysis. The nature of this analysis depends on which neural notions have been investigated. For experiments that make use of neural notions that are similar to contrastive analysis, a simple statistical procedure is available, that can be applied out of the box. We review this procedure in Section 11.4.1. Section 11.4.2 describes how to proceed in other cases.

11.4.1. *Ready-to-Use Statistics.* CoAA can be applied to a wide range of neural states and neural systems, which is why one can apply CoAA to a host of new experimental possibilities. However, one doesn't have to. CoAA is particularly easy to apply to the neural notions that are already of use in contrastive analysis, and doing so brings a number of advantages, for example:

- improved results, as compared to contrastive analysis (Theorem 19).
- better error terms (Proposition 21, cf. below).
- the possibility of using more and/or different states of consciousness.

In Section 7.1, we have derived a statistical procedure precisely for this case. Starting from the mathematics of CoAA developed in earlier sections, and working with the case of an fMRI analysis for simplicity, we derived a statistical procedure that defines a methodology to search for NCCs in this case. This procedure is "ready-to-use"; no further mathematical work is required for an application. It can be applied as follows:

1. **Measurements.** Measure fMRI activity as usual, and apply a suitable measure of consciousness to infer which state of consciousness, out of the set of states of consciousness that have been chosen for the analysis, is active during a scan. We denote the activity values of the different voxels of an fMRI scan by Y_i.
2. **Run MANCOVA Tests.** For each pair e, e' of distinct states of consciousness in the set of states of consciousness, and for each of the ROIs S under consideration, run a MANCOVA test with the following settings:
   - dependent variables are the activity values Y_i of each voxel i of an ROI S.
   - the group variable comprises the two states of consciousness.
   - covariates are as collected in the experiment.
3. **Analyse MANCOVA Tests.** For each ROI S, check if the MANCOVAs of this ROI for all pairs of states of consciousness are significant.
4. **Find NCCs.** According to Theorem 20, the minimal ROI for which all MANCOVA tests are significant *is an NCC* as defined by (Chalmers, 2000) and (Crick & Koch, 1990). Furthermore, every ROI for which it is not the case that all MANCOVAs are significant, or which is not minimal in this class, is not an NCC as defined by (Chalmers, 2000) and (Crick & Koch, 1990).

Details of this procedure are explained in Section 7.1, and adaptations to measurement techniques other than fMRI should be straightforward.

In Proposition 21, we have derived α and β error rates for the above procedure. Denoting the α and β error rates of the individual MANCOVAs by α_0 and β_0, to first order in α_0 and

β_0, the overall error rates of this procedure are given by:

$$\alpha \leq J{\cdot}\alpha_0 + L{\cdot}\beta_0$$
$$\beta \leq \max(J{\cdot}\alpha_0, \beta_0)\,,$$

where J is the number of distinct pairs of states of consciousness used in the experiment, and where L is the number of ROIs under consideration that are inside of the NCC.[22]

These error rates might be of interest for purely statistical reason, as both J and L can be expected to be small single digit numbers. For example, for three states of consciousness, we have $J = 3$. Furthermore, dependencies that one would otherwise expect, e.g. on the number of voxels in an fMRI analysis, do not exist. Correspondingly, an application of CoAA might not suffer from the problem of multiple comparisons as usual the case for fMRI studies, which might render the need for cluster-based analyses, for example as proposed in (Heller, Stanley, Yekutieli, Rubin, & Benjamini, 2006), obsolete.[23]

11.4.2. *General CoAA.* The statistics explained in the last section presume that the neural systems and neural states that are to be investigated—the first and second elements of the analysandum explained in Section 11.1 above—are, essentially, those that are used in contrastive analysis. While some variations are possible, e.g. regarding the choice of neural system, these options are rather limited.

There is a host of reasons of why one might want to search for NCCs among different neural systems, some of which we have discussed in Section 9 above.

One advantage of the general mathematical framework of CoAA is that CoAA is not premised on a single choice of neural system or neural states. CoAA can be applied to *any choice* of neural system and neural state that can be measured in an experiment.

Here, 'measured' can be interpreted very widely. All that is required for an application of CoAA is information about which neural states were active in experimental trials (together with which of the states of consciousness under consideration). This information constitutes the co-activation data that we have introduced in Section 3.

It is likely that many experimental protocols deliver this information right out of the box, in which case CoAA can be applied without the need of further theoretical work. If this is not the case, for example due to a larger deductive distance between the neural notions of interest and the output of a neuroimaging device, one can use a statistical procedure to infer which neural states were active during experimental trials.

CoAA does not place restrictions on how information about the neural states is obtained in an experiment; any neuroimaging tool, and any statistical procedure goes. In Section 7.2, we briefly explain one class of procedures that might be used. This class can be used if a likelihood function that specifies the expected neuroimaging data for the neural states under consideration is available. But nothing hinges on this particular proposal, any other statistic can be used as well.

Once data about the co-activation of neural states and states of consciousness is available, the analysis of NCCs proceeds in terms of three formal objects. These formal objects give a computation pipeline which should be easy to implement in practice. Writing 'A $\twoheadrightarrow$ B' to indicate that B can be computed once A is available, this computation pipeline is:

$$\mathbf{D} \twoheadrightarrow \mathbf{B}_N(e) \twoheadrightarrow \mathbf{S}_N(e) \twoheadrightarrow \mathbf{M} \twoheadrightarrow \mathbf{NCC}\,.$$

Here, $\mathbf{D}$ is the empirical data. The sets $\mathbf{B}_N(e)$ are defined at the beginning of Section 5; they contain those neural states of a system N that have been measured as co-active with a state of consciousness e. The sets $\mathbf{S}_N(e)$ comprise neural states of N that meet the sufficiency requirement in Definition 1; they can be computed based on $\mathbf{B}_N(e)$ as described in Lemma 5, The set M comprises neural systems that meet the mapping requirement in Definition 1; they can be computed based on the sets $\mathbf{S}_N(e)$ as described in Definition 8. The NCCs, finally can be computed based on Theorem 14. The theorem states that if a system N is a minimal element of $\mathbf{M}$, and if the system satisfies a

[22] Explicit definitions of J and L are provided in Footnote 13.

[23] I am grateful to Maximilian Kathofer for discussions on cluster analysis and related topics.

condition which we discuss next, then by mathematical necessity, N is an NCC as defined by (Chalmers, 2000) and (Crick & Koch, 1990).

The condition that Theorem 14 makes use of, in addition to the minimality of N in **M**, is that N is of full measure in **D** (Definition 3). For most statistics, this condition amounts to a Type II error control (β error rate) for tests that identify the states of the system N.

11.5. **Analysis of Existing Data.** A final option that should probably be mentioned in the context of this overview concerns the possibility of re-analysing existing data.

As far as experimental presumptions are concerned, CoAA is, in a strict sense, a *generalization* of contrastive analysis. This means that, mathematically speaking, if an experiment provides data that can be analysed with contrastive analysis, then this data can also be analysed with CoAA. Furthermore, as far as the logic of measurement is concerned, CoAA can be shown to *improve upon* the result of contrastive analysis (cf. Theorem 19). This raises the question of whether one should attempt to re-analyse existing data with contrastive analysis.

From a mathematical perspective, the answer is clearly yes. The procedure of Section 11.4.1 can, likely, be applied to many of the data sets that exist already. And in particular in cases where only two states of consciousness are present, this procedure constitutes a simple and statistically powerful empirical test for NCCs.

At the present stage, it is completely open what such re-analysis would give. Given the differences in underlying logic, it is very unlikely that a CoAA-based re-analysis of existing data would identify the exact same activity patterns that contrastive analysis found. It is likely that a subset would be found, but it is also possible that CoAA finds no NCC at all, for example if the measured activity is due to confounders (Section 8).

12. Conclusion

The search for Neural Correlates of Consciousness (NCCs) is a demanding empirical endeavour, requiring the interplay of experience and expertise across experimental, theoretical, and philosophical domains.

In this paper, we attempt a contribution from a novel domain, that of mathematics. The initial goal of this project was to analyse the logic of measuring NCCs. Here, 'logic' refers to the logical connections between concepts, definitions, and mathematical tools, as identified by lemmas, theorems, and propositions. The project was aimed at answering questions like: What relation, exactly, is there between the canonical definition and the type of data that is presumed in contrastive analysis? Under which circumstances, exactly, can the NCC as canonically defined be measured empirically? What can be known in case of incomplete data? What exactly is it that contrastive analysis measures within the canonical definition?

As the analysis of these questions made progress, however, a somewhat surprising implication emerged. Could it be the case that what the logic is pointing at is a new methodology to measure NCCs in the lab? Perhaps something that can complement contrastive analysis and decoding in empirical practice?

As a result of this possibility, the nature of the project shifted. The goal, then, became to understand and explore this new methodology.

A first realisation was that the new methodology might be more general than the contrastive approach. There was a first theorem that showed that, as far as the logic of measurement is concerned, the methodology seemed to improve upon results of contrastive analysis (Theorem 19). A second realisation was that the methodology can be applied to neural states and states of consciousness that might not be particularly well-suited for contrastive analysis (Sections 9 and 10).

Then came another surprise: that if applied to the setting of contrastive analysis, the method gives rise to a simple statistical procedure (Section 7, Theorem 20) that could be easily applied in experiments, and perhaps even to existing data. Furthermore, the derivation of α and β error rates for this statistic (in Section 7, Proposition 21) identified peculiar statistical advantages; for example, that the error rates do not depend on the number of voxels in an fMRI analysis, which might be relevant

for the multiple testing problem in an fMRI-based search for NCCs. Finally, it became apparent that the method might also have implications for the notorious problem of confounders in NCC research (Section 8).

With each of these discoveries, there came a bit more hope that this methodology might actually be something that could support empirical research in the field; something that isn't just scratching the itch of a mathematician's curiosity, but that, perhaps, could be of actual help. But at the same time, of course, it was painfully clear that the actual litmus test for this work will come only once the mathematical part has been completed, when it is presented to the experimental expertise in the field at large. It is in this spirit that the paper is submitted to the field.

If the mathematical perspective on NCCs provided here turned out to be helpful to the field, a whole range of further research could be envisaged. For example:

1. a mathematical analysis of the logic of the decoding-based search for NCCs, as developed by Haynes (2009).
2. exploration of the application of structural information in the search for NCCs, as already suggested by Chalmers (2000), and further explored, for example, in (Fink et al., 2021) and (Kleiner, 2024).
3. the development of statistical tools tailored to specific research programmes, for example those proposed in (Cleeremans, 2005; Hohwy & Seth, 2020; Mudrik et al., 2024; Luppi et al., 2021; Lyre, 2022; Cecconi, van der Lande, & Sala, 2024).
4. the development of statistical tools for alternative definitions of NCCs or CCCs, such as those proposed by (Fink, 2016; Wiese & Friston, 2021; Mckilliam, 2024; Cleeremans, 2005).
5. a mathematical analysis of how experimental data from different sources and potentially across different paradigms could be combined in a collaborative search for the NCC.

In summary, this paper is a perhaps somewhat unusual reaction to Lepauvre and Melloni (2021)'s recent conjecture that "[d]espite the clear advantages of the contrastive method, its mostly exclusive use in the research of consciousness may have constrained the field too much, becoming an obstacle to finding NCCs" (p. 13). The methodology discovered here, preliminarily called *Co-Activation Analysis* (or *CoAA* for short), is a mathematical contribution at "broadening the experimental methods [to] help us (...) better understand the phenomenon" of consciousness (ibid.).

Acknowledgments. I would like to thank Rony Hirschhorn, Pedro Mediano, Jonathan Mason, Maximilian Kathofer, Tim Ludwig, and David Chalmers, as well as the participants of the 2025 winter school of the Mediterranean Society for Consciousness Science (MESEC), for valuable discussions on NCC research. Furthermore, I would like to thank Jonathan Mason for valuable feedback on an early version of the manuscript.

References

Andrillon, T. (2023). How we sleep: from brain states to processes. *Revue Neurologique*, *179*(7), 649–657.

Aru, J., Axmacher, N., Do Lam, A. T., Fell, J., Elger, C. E., Singer, W., & Melloni, L. (2012). Local category-specific gamma band responses in the visual cortex do not reflect conscious perception. *Journal of Neuroscience*, *32*(43), 14909–14914.

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(2), 737–746.

Baars, B. J. (1986). What is a theory of consciousness a theory of?–The search for criterial constraints on theory. *Imagination, Cognition and Personality*, *6*(1), 3–23.

Bachmann, T. (2009). Finding erp-signatures of target awareness: puzzle persists because of experimental co-variation of the objective and subjective variables. *Consciousness and cognition*, *18*(3), 804–808.

Bachmann, T., & Hudetz, A. G. (2014). It is time to combine the two main traditions in the research on the neural correlates of consciousness: C= L× D. *Frontiers in Psychology*, *5*, 940.

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness? *Trends in cognitive sciences*, *20*(6), 405–413.

Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences*, *23*(12), 1003–1013.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Cecconi, B., Bonhomme, V., Laureys, S., Gosseries, O., Boly, M., & Annen, J. (2025). Experimental approaches to study sensory disconnection in humans during sleep and anesthesia. *Current Opinion in Behavioral Sciences*, *63*, 101505.

Cecconi, B., van der Lande, G., & Sala, A. (2024). Neural correlates of consciousness. In *Coma and disorders of consciousness* (pp. 1–15). Springer.

Chalmers, D. J. (2000). Neural correlates of consciousness: Empirical and conceptual questions. In T. Metzinger (Ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (pp. 17–39). MIT Press.

Cleeremans, A. (2005). Computational correlates of consciousness. *Progress in brain research*, *150*, 81–98.

Cohen, M. A., Ortego, K., Kyroudis, A., & Pitts, M. (2020). Distinguishing the neural correlates of perceptual awareness and postperceptual processing. *Journal of Neuroscience*, *40*(25), 4925–4935.

Crick, F. (1994). *Astonishing Hypothesis: The Scientific Search for the Soul*. Simon and Schuster.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences* (Vol. 2, pp. 263–275).

De Graaf, T. A., Hsieh, P.-J., & Sack, A. T. (2012). The 'correlates' in neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(1), 191–197.

Dellert, T., Balster, H., Schlossmacher, I., Bruchmann, M., Moeck, R., & Straube, T. (2025). Neural correlates of consciousness in an auditory no-report fmri study. *bioRxiv*. doi: 10.1101/2025.05.16.654468

Dellert, T., Müller-Bardorff, M., Schlossmacher, I., Pitts, M., Hofmann, D., Bruchmann, M., & Straube, T. (2021). Dissociating the neural correlates of consciousness and task relevance in face perception using simultaneous EEG-fMRI. *Journal of Neuroscience*, *41*(37), 7864–7875.

Demertzi, A. (2023). Neural correlates of mind blanking in the human brain: Challenges for consciousness theories. In *6th panhellenic conference of cognitive science.*

Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., ... others (2019). Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Science advances*, *5*(2), eaat7603.

Demsar, E., & Windt, J. (2024). Studying dream experience through dream reports: Points of contact between dream research and first-person methods in consciousness science. In *Dreaming and memory: Philosophical issues* (pp. 85–117). Springer.

Duman, I., Ehmann, I. S., Gonsalves, A. R., Gültekin, Z., Van den Berckt, J., & van Leeuwen, C. (2022). The no-report paradigm: a revolution in consciousness research? *Frontiers in Human Neuroscience*, *16*, 861517.

Fink, S. B. (2016). A deeper look at the "neural correlate of consciousness". *Frontiers in Psychology*, *7*, 1044.

Fink, S. B. (2020). A double anniversary for the neural correlates of consciousness: Editorial introduction. *Philosophy and the Mind Sciences*, *1*(II).

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443.

Förster, J., Koivisto, M., & Revonsuo, A. (2020). ERP and MEG correlates of visual consciousness: The second decade. *Consciousness and Cognition*, *80*, 102917.

Frässle, S., Sommer, J., Jansen, A., Naber, M., & Einhäuser, W. (2014). Binocular rivalry: frontal activity relates to introspection and action but not to perception. *Journal of Neuroscience*, *34*(5), 1738–1747.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, *29*(1), 1–49.

Gurney, E. (1881). *Monism. Mind, VI (22), 153–173.*

Hatamimajoumerd, E., Murty, N. A. R., Pitts, M., & Cohen, M. A. (2022). Decoding perceptual awareness across the brain with a no-report fMRI masking paradigm. *Current Biology*, *32*(19), 4139–4149.

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in cognitive sciences*, *13*(5), 194–202.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., & Benjamini, Y. (2006). Cluster-based analysis of FMRI data. *NeuroImage*, *33*(2), 599–608.

Hohwy, J. (2009). The neural correlates of consciousness: new experimental approaches needed? *Consciousness and cognition*, *18*(2), 428–438.

Hohwy, J., & Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II).

Hudetz, A. G. (2024). Does consciousness have dimensions? *Journal of Consciousness Studies*, *31*(7-8), 55–73.

Irvine, E. (2013). Measures of consciousness. *Philosophy Compass*, *8*(3), 285–297.

Jang, H., Mashour, G. A., Hudetz, A. G., & Huang, Z. (2024). Measuring the dynamic balance of integration and segregation underlying consciousness, anesthesia, and sleep in humans. *Nature communications*, *15*(1), 9164.

Klein, C., Hohwy, J., & Bayne, T. (2020). Explanation in the science of consciousness: From the neural correlates of consciousness (nccs) to the difference makers of consciousness (dmcs). *Philosophy and the Mind Sciences*, *1*(II).

Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, *119*.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, *93*(3), 480–490.

Kriegel, U. (2020). Beyond the neural correlates of consciousness. *The Oxford handbook of the philosophy of consciousness*, 261–276.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, *103*(49), 18763–18768.

Lee, A. Y. (2023). Degrees of consciousness. *Noûs*, *57*(3), 553–575.

Lepauvre, A., & Melloni, L. (2021). The search for the neural correlate of consciousness: Progress and challenges. *Philosophy and the Mind Sciences*, *2*.

Luppi, A. I., Cain, J., Spindler, L. R., Górska, U. J., Toker, D., Hudson, A. E., ... others (2021). Mechanisms underlying disorders of consciousness: bridging gaps to move toward an integrated translational science. *Neurocritical care*, *35*, 37–54.

Lyre, H. (2022). Neurophenomenal Structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, *2022*(1), niac012.

Marshall, H. R. (1901). Consciousness, self-consciousness and the self. *Mind*, *10*(37), 98–113.

Marvan, T., & Polák, M. (2020). Generality and content-specificity in the study of the neural correlates of perceptual consciousness. *Philosophy and the Mind Sciences*, *1*(II).

Matthews, J., Schröder, P., Kaunitz, L., Van Boxtel, J. J., & Tsuchiya, N. (2018). Conscious access in the near absence of attention: critical extensions on the dual-task paradigm. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170352.

Mazor, M., & Fleming, S. M. (2020). Distinguishing absence of awareness from awareness of absence. *Philosophy and the Mind Sciences*, *1*(II).

Mckilliam, A. (2020). What is a global state of consciousness? *Philosophy and the Mind Sciences*, *1*(II).

Mckilliam, A. (2024). A mechanistic alternative to minimal sufficiency as the guiding principle for ncc research. *Neuroscience of Consciousness*, *2024*(1), niae014.

Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *Journal of Neuroscience*, *31*(4), 1386–1396.

Melloni, L., & Singer, W. (2010). Distinct characteristics of conscious experience are met by large-scale neuronal synchronization. In *New horizons in the neuroscience of consciousness* (pp. 17–28). John Benjamins Publishing Company.

Metzinger, T. (2000). *Neural correlates of consciousness: Empirical and conceptual questions*. MIT Press.

Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of "pure" consciousness. *Philosophy and the Mind Sciences*, *1*(I), 1–44.

Metzinger, T. (2024). *The elephant and the blind: the experience of pure consciousness: philosophy, science, and 500+ experiential reports*. MIT Press.

Metzner, C., Schilling, A., Traxdorf, M., Schulze, H., & Krauss, P. (2021). Sleep as a random walk: A super-statistical analysis of eeg data across sleep stages. *Communications Biology*, *4*(1), 1385.

Michel, M. (2019). The mismeasure of consciousness: A problem of coordination for the perceptual awareness scale. *Philosophy of Science*, *86*(5), 1239–1249.

Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *14*(2), e1628.

Mormann, F., & Koch, C. (2007). Neural correlates of consciousness. *Scholarpedia*, *2*(12), 1740.

Mudrik, L., Hirschhorn, R., & Korisky, U. (2024). Taking consciousness for real: Increasing the ecological validity of the study of conscious vs. unconscious processes. *Neuron*.

Nani, A., Manuello, J., Mancuso, L., Liloia, D., Costa, T., & Cauda, F. (2019). The neural correlates of consciousness and attention: two sister processes of the brain. *Frontiers in Neuroscience*, *13*, 1169.

Overgaard, M. (2004). Confounding factors in contrastive analysis. *Synthese*, *141*, 217–231.

Overgaard, M. (2015). The challenge of measuring consciousness. *Behavioral Methods in Consciousness Research*, 7–20.

Overgaard, M., & Fazekas, P. (2016). Can no-report paradigms extract true correlates of consciousness? *Trends in cognitive sciences*, *20*(4), 241–242.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Paßler, M. (2023). The exclusionary approach to consciousness. *Neuroscience of Consciousness*, *2023*(1), niad022.

Pauen, M., & Haynes, J.-D. (2021). Measuring the mental. *Consciousness and Cognition*, *90*, 103106.

Peters, M. A., Kentridge, R. W., Phillips, I., & Block, N. (2017). Does unconscious perception really exist? continuing the assc20 debate. *Neuroscience of consciousness*, *2017*(1), nix015.

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive sciences*, *5*(3), 229–269.

Petitmengin, C., Remillieux, A., & Valenzuela-Moguillansky, C. (2019). Discovering the structures of lived experience: Towards a micro-phenomenological analysis method. *Phenomenology and the Cognitive Sciences*, *18*(4), 691–730.

Phillips, I. B., & Morales, J. (2020). The fundamental problem with no-cognition paradigms. *Trends in Cognitive Sciences*.

Ramstead, M. J., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., ... others (2022). From generative models to generative passages: a computational approach to (neuro) phenomenology. *Review of Philosophy and Psychology*, *13*(4), 829–857.

Reggia, J. A., Katz, G. E., & Davis, G. P. (2019). Modeling working memory to identify computational correlates of consciousness. *Open Philosophy*, *2*(1), 252–269.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in cognitive sciences*, *12*(8), 314–321.

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., ... Tononi, G. (2017). The neural correlates of dreaming. *Nature neuroscience*, *20*(6), 872–878.

Siclari, F., Patriota, J., & Olcese, U. (2025). Dreaming as a window on the mechanisms of consciousness. PsychArchives.

Singer, W. (2014). The ongoing search for the neuronal correlate of consciousness. In *Open mind*. Open MIND. Frankfurt am Main: MIND Group.

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of mathematical psychology*, *107*, 102632.

Sparby, T., Leass, M., Weger, U. W., & Edelhäuser, F. (2023). Training naive subjects in using micro-phenomenological self-inquiry to investigate pain and suffering during headaches. *Scandinavian Journal of Psychology*, *64*(1), 60–70.

Stevner, A., Vidaurre, D., Cabral, J., Rapuano, K., Nielsen, S. F. V., Tagliazucchi, E., ... others

(2019). Discovery of key whole-brain transitions and dynamics during human wakefulness and non-REM sleep. *Nature communications*, *10*(1), 1–14.

Tal, H., Wright, M., Prest, S., Sandved-Smith, L., & Sacchet, M. (2025). Active inference, computational phenomenology, and advanced meditation: Toward the formalization of the experience of meditation. *Preprint*.

Tononi, G., Boly, M., & Cirelli, C. (2024). Consciousness and sleep. *Neuron*, *112*(10), 1568–1594.

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: an update. *Annals of the New York Academy of Sciences*, *1124*(1), 239–261.

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. (2015). No-report paradigms: extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770.

Tull, S., Kleiner, J., & Smithe, T. S. C. (2023). Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy. *arXiv preprint arXiv:2308.00861*.

Ward, J. (1911). Psychology. In *In Encyclopedia britannica* (Vol. XXII, p. 547–604). Cambridge University Press.

Wiese, W., & Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *Philosophy and the Mind Sciences*, *2*.