Inference to the Best Neuroscientific Explanation

Davide Coraci, Gustavo Cevolani, Igor Douven

Inference to the Best Neuroscientific Explanation

Davide Coraci[1], Gustavo Cevolani[1], and Igor Douven[2]

[1]IMT School for Advanced Studies Lucca
[2]IHPST/CNRS/Panthéon–Sorbonne University

December 31, 2024

Abstract

Neuroscientists routinely use reverse inference (RI) to draw conclusions about cognitive processes from neural activation data. However, despite its widespread use, the methodological status of RI is a matter of ongoing controversy, with some critics arguing that it should be rejected wholesale on the grounds that it instantiates a deductively invalid argument form. In response to these critiques, some have proposed to conceive of RI as a form of abduction or inference to the best explanation (IBE). We side with this response but at the same time argue that a defense of RI requires more than identifying it as a form of IBE. In this paper, we give an analysis of what determines the quality of an RI conceived as an IBE and on that basis argue that whether an RI is warranted needs to be decided on a case-by-case basis. Support for our argument will come from a detailed methodological discussion of RI in cognitive neuroscience in light of what the recent literature on IBE has identified as the main quality indicators for IBEs.

Keywords: abduction; explanation; functional magnetic resonance imaging; inference to the best explanation; neuroscience; reverse inference.

1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is one of the main neuroscientific techniques for studying associations between neural evidence (i.e., the observed activation of certain brain regions) and cognitive hypotheses (i.e., hypotheses about the recruitment of certain cognitive processes during some task). Practitioners of this technique tend to rely on a mode of reasoning known as "reverse inference" (RI for short; see Poldrack 2006). Specifically, RI is used to draw conclusions about the engagement of a certain cognitive process or function from the registered activation of a certain brain region. Despite its prevalent use, the methodological status of RI is a subject of ongoing debate (for critical discussion, see Poldrack 2006, 2011; Bourgeois-Gironde 2010; Hutzler 2014; Machery 2014; Glymour and Hanson 2016; Nathan and Del Pinal 2017; Coraci, Calzavarini, and Cevolani 2022), with some critics proposing that neuroscientists abstain entirely from using it (e.g., Anderson 2010).

In an effort to validate its use, some scholars have posited RI as a form of abductive reasoning or inference to the best explanation (IBE for short; see Bourgeois-Gironde 2010; Poldrack 2011; Calzavarini and Cevolani 2022), which, although not deductively valid, has been heralded as the cornerstone of scientific methodology (e.g., Boyd 1984; McMullin 1992). However, simply acknowledging RIs as IBEs is not enough to justify their use in neuroscience, given that even

some of the staunchest advocates of IBE have argued that we are licensed to use it only if certain conditions are met (e.g., Lipton 1993; Bird 2010; Douven 2022).

In this paper, we side with previous work which proposes to view RI as a form of IBE. However, we go beyond that work by presenting a more nuanced and qualified defense of RI as IBE, most notably by giving an accurate analysis of what determines the quality of an RI conceived as IBE. Doing so will help us argue that whether an RI is warranted needs to be decided on a case-by-case basis and also is not always a matter of *yes* or *no*; rather, RIs can be warranted more or less strongly. Support for our argument will come from a detailed methodological discussion of RI in cognitive neuroscience in light of what the recent literature on IBE has identified as the main conditions for an IBE to be warranted. The discussion will involve some case studies from the neuroscientific literature which jointly show that whether, and if so to what extent, RI satisfies those conditions crucially depends on the specific method or methods it relies on.

Section 2 delineates RI in cognitive neuroscience, underscoring the inadequacies of current discussions of RI as an IBE. Section 3 compares the major methods used in implementing RIs based on fMRI data, with a distinct focus on the differences between univariate and multivariate methods. Here, we place special emphasis on Representational Similarity Analysis (Kriegeskorte 2008), a multivariate technique that has so far received little attention in the RI debate. Section 4 reviews recent theoretical and empirical work on IBE, highlighting several principles suggested by philosophers for evaluating the quality of competing explanations and, more generally, the validity of an abductive inference. Section 5 illustrates how these principles can advance methodological discussions surrounding RI in cognitive neuroscience.

2 The Status of Reverse Inference

According to Poldrack (2006), the most common version of RI has the following structure:

P1. In the literature, when cognitive process *Cog* was engaged (during task *T*), then brain area *Act* was active.
P2. In the present study, brain area *Act* was active.
C. The activity of area *Act* in the present study demonstrates the engagement of cognitive process *Cog* (during task *T*).

Thus schematized, RI appears to instantiate a form of reasoning generally referred to as "affirming the consequent" (AC):

P1. If *Cog* then *Act*
P2. *Act*
C. *Cog*

AC is known to be a fallacy. If it rains the streets will be wet, but it would be wrong to infer from the streets being wet that it has rained: the streets could be wet for any number of reasons. Specifically in relation to RI, this problem has been referred to as "the problem of selectivity of brain regions" (Poldrack 2006; Nathan and Del Pinal 2017), which is the fact that, often, the activity of a region *Act* can be associated with a number of different cognitive functions rather than just with a single cognitive function *Cog*. Due to this "multifunctionality" of brain areas, which is crucially acknowledged in neuroscience, inferring from *Act* to a specific cognitive function *Cog* is always risky, since *Act* could currently contribute to a another function different

from *Cog*.[1] As a result, there is much scepticism among researchers about reasoning from brain activations to cognitive functions via RI (e.g., Anderson 2010; Bourgeois-Gironde 2010; Poldrack 2011; Glymour and Hanson 2016; Nathan and Del Pinal 2017; Calzavarini and Cevolani 2022).

More optimistic proposals conceptualize RI as a fallible but still useful form of inference. Some of these view RI as an instance of probabilistic or inductive reasoning, along Bayesian (Poldrack 2006; Hutzler 2014) or likelihoodist (Machery 2014) lines. But, as previously mentioned, RI has also been analyzed as a form of IBE, that is, roughly, as an inference from observed evidence to putative explanations or causes (Poldrack 2011; Glymour and Hanson 2016; Calzavarini and Cevolani 2022).[2] From these more optimistic perspectives, RI appears as an inferential strategy useful for figuring out, and even confirming one of, a number of competing neuroscientific hypotheses able to account for a given body of fMRI evidence.

Interestingly, such proposals moved from conceptualizing RI as an instance of abductive reasoning in Peirce's sense—so as having a place in the context of discovery (e.g., Deeley 1994)—to viewing RI as an instance of IBE, so as having a place in the context of justification. For instance, Poldrack (2006) was the first to link RI to abduction as originally conceived by Peirce, noticing that RI may be a useful strategy for discovering new hypotheses concerning the recruitment of certain cognitive processes on the basis of the observed neural activation. In line with this view, Bourgeois-Gironde (2010) analyzes RI as a form of abduction leading to the formulation of tentative hypotheses in need of further testing. Similarly, Glymour and Hanson (2016) argue that the real problem of RI "is not confirmation but search: how to find among the huge number of alternatives the hypothesis, or hypotheses, that best explain the data" (p. 1150), even if this statement can be read as to point both to a discovery or a justification role for RI. In general however, none of these authors provide a definitive word about the status of RI. Rather, they open the way for further discussion of RI not only within the context of discovery but also within that of the justification of cognitive hypotheses. Thus, the debate appears to follow the same path of the philosophical discussion that, historically, concerned the interpretation of the notion of abduction, with some authors pointing toward its exploratory component and others stressing its justificatory role as IBE.

In this connection, Calzavarini and Cevolani (2022) represents the first attempt to systematically discuss RI with respect to the philosophical notion of abductive inference, suggesting to distinguish between a weak and a strong version of RI in order to disentangle the "heuristic" or "discovery" role of RI and its "justificatory" or "confirmatory" role as a form of IBE. As these authors show, this distinction is important to properly understand the current debate about RI, and especially to appreciate the role and pervasiveness of weak RI in the neuroscientific literature. As far as strong RI is concerned, however, Calzavarini and Cevolani (2022) do not go beyond discussing an (interesting) case study; in particular, they do not provide any general analysis of how to assess the reliability of RI as IBE.[3]

[1] This motivates what Burnston (2016) calls "contextualism" in connection to the localization of cognitive functions in brain areas.

[2] This might seem unhelpful, given that IBE has itself been said to instantiate AC. That is a mistake, however, for more than one reason (as will be seen), but one obvious reason is that, at a minimum, one would want to add a premise to the schema given in the main text to the effect that *Cog* itself, and also the connection between *Cog* and *Act*, is to satisfy certain criteria for the inference to go through. The resulting schema would no longer be the one used to characterize AC.

[3] The same is true of the recent overview of RI offered by Coraci, Calzavarini, and Cevolani (2022). In Section 5.2, we will provide a critical reassessment of the case study discussed by Calzavarini and Cevolani (2022), by specifically

Therefore, up to this point, the debate about the *status* of RI misses a detailed and qualified analysis of RI as a form of IBE. For reasons to be given in Section 4, we hold that inferring to the best explanation can be warranted, depending on whether certain conditions or desiderata are met. Accordingly, we argue for a qualified rather than a blanket endorsement of RI, meaning that, in our view, there is no *general* answer to the question of whether, and if so, to which extent, we are warranted in accepting a cognitive hypothesis on the basis of neural evidence. The reason for this is that much depends on the specific (statistical and experimental) methods applied to perform RIs, as well as on the details of concrete cases and studies. Even so, it is still possible to derive from the literature on IBE general guidelines for how to decide the said question in a given context. In Section 5, we illustrate the usefulness of these guidelines by discussing and comparing different applications of RI. To this purpose, we first need to see in some detail how RI is routinely performed in cognitive neuroscience, in order to appreciate the extent to which the quality of the provided explanation depends on the specific method used to analyze fMRI data.

3 From Location-based Reverse Inference to Representational Similarity Analysis

Current research in neuroscience employs various forms of analysis even among experimental studies based on fMRI. A relevant distinction, which is the focus of this section, depends on whether "univariate" or "multivariate" statistical techniques are applied to analyze the available fMRI data. This leads to two different forms of RI, to wit, "location-based" versus "pattern-based" reverse inference (following Del Pinal and Nathan 2017; Nathan and Del Pinal 2017). Here, we briefly present the two methods, paying special attention to one prominent multivariate technique, viz., Representational Similarity Analysis. Further on, this will help us assess the impact of the various methods on the quality of the inferred cognitive explanations and discuss whether and to what extent those explanations meet the guidelines suggested in the literature on IBE.

3.1 Location-based Reverse Inference

The more conventional, and still more widely used, form of RI is the one we may call "location-based" or "voxel-based RI," since it is performed via a univariate statistical analysis of fMRI data concerning individual voxels.[4] To get an idea of how it works, suppose we are interested in locating the brain areas that show higher activation during a specific experimental manipulation (e.g., a task involving face perception). The initial scanning stage consists in acquiring the neural signal from the same participants both under the experimental manipulation and at rest (or, more generally, during a contrast condition). Then, a statistical analysis is performed on the acquired neural signal to compare the average neural activity recorded during the two conditions. Those brain areas that show a statistically significant difference will be interpreted as having been more active due to the manipulation. When many voxels report a positive and significant change in their neural activity, these are gathered together in clusters or regions,

focusing on the quality of the RI drawn in that study (Lieberman and Eisenberger 2015) in the light of the philosophical literature about IBE.

[4] A voxel can be thought of as a three-dimensional pixel and represents the minimal unit at which the MRI scanner can record the neural signal. The size of the voxel, usually between 1 and 4 millimeters, defines the resolution of the brain image. Even in high-resolution imaging, a single voxel includes thousands of neurons.

and inferences about the engagement of the cognitive function at issue are drawn at the cluster level.

While the majority of currently published fMRI studies rely on this type of data analysis, various criticisms have been raised recently against location-based analysis, and especially against RI based on this technique (see especially Davis, LaRocque, et al. 2014; Del Pinal and Nathan 2017; Nathan and Del Pinal 2017). One point of critique is that this kind of RI relies on a prior assumption about the neural localizability of cognitive processes, which is problematic in light of the lack of selectivity of brain regions mentioned previously (Del Pinal and Nathan 2017). A second point of critique concerns the association between the cognitive process engaged in the experimental task and the observed activation, which, according to the critics, is difficult to establish in a precise and quantitative manner. Indeed, in many univariate studies, the "association laws" between cognitive processes and brain regions appears to rest on potentially biased and unsystematic reviews of the available literature (Nathan and Del Pinal 2017, p. 10 f).[5] Finally, one of the main limitations of univariate methods is that neuropsychological variables putatively elicited during experimental manipulations are fully coded by the activity of a single voxel (e.g., Davis, LaRocque, et al. 2014). This assumption makes the mapping of potentially multidimensional characteristics of experimental stimuli much harder, in particular when the investigated cognitive function appears so complex and multifaceted that it requires the analysis of a combination of neural information coming from multiple loci.

3.2 Pattern-based Reverse Inference

The issues with standard, location-based analysis mentioned above motivated the development of other methods, leading to what may be called "pattern-based" or "pattern-decoding" RI. This second form of RI is based on so-called multivariate pattern analysis (MVPA) of fMRI data.[6] Introduced in a seminal work by Haxby and colleagues (Haxby, Gobbini, et al. 2001), MVPA has received increasing attention in the fMRI community, since it allows researchers to investigate how populations of voxels, rather than single units, encode information. MVPA encompasses different steps, which it is useful to describe in some detail to understand how this method differs from standard univariate analysis (Norman et al. 2006; Haynes and Rees 2006).

After fMRI data acquisition, MVPA starts by selecting a subset of the measured signals, for instance, a specific region of the brain corresponding to a three-dimensional grid of N voxels. Next, through so-called pattern assembly (Norman et al. 2006), these neuroimaging data are sorted into pattern vectors reporting the change in the neural activity for each of the selected voxels over the time course of the experiment. Such vectors can be represented in an N-dimensional feature space and are labeled according to the experimental conditions that

[5] The tools and software for conducting large-scale, automated syntheses of the fMRI literature that have become available to neuroscientists over the past ten years (see, e.g., Yarkoni et al. 2011; Poldrack 2011; Costa et al. 2021) have somewhat helped to undercut Nathan and Del Pinal's criticism. These authors are aware of this development, but they claim that—at least at the time of their writing, viz., 2017—machine-learning decoding methods that would provide the automated syntheses of the fMRI literature, as discussed for instance in Poldrack (2011), are "still largely ignored in critical and methodological discussions of reverse inference" (Nathan and Del Pinal 2017, p. 8 f).

[6] The terminology is not entirely settled in the literature. For a discussion concerning the use of the partially overlapping and interconnected concepts of "multivariate analysis," "multi-voxels analysis," "pattern analysis," "mind-reading," and the "encoding/decoding" distinction, see Kriegeskorte and Bandettini (2007) and Ritchie, Kaplan, and Klein (2020). For a technical introduction to MVPA, see Haynes and Rees (2006), Norman et al. (2006), Davis and Poldrack (2013), and Haxby, Connolly, and Guntupalli (2014).

generated the pattern. Then, the data pattern vectors are split into two sets, the training set and the test set.

At this point, a statistical classifier (e.g., a machine learning algorithm able to automatically classify data into clusters) is applied in order to find an optimal partition of the neural data within the *N*-dimensional space, according to which experimental conditions can be effectively discriminated.[7] The training set is used to feed the algorithm for the classification: if the classifier's algorithm is able to distinguish between different populations of voxels based on their activity, such populations can be interpreted as responding to the experimental conditions with which they have been previously labeled. The trained classifier is then validated on the test set in order to assess its accuracy in predicting the experimental conditions when "new," unseen neural data are taken as input. The performance is usually assessed through a cross-validation procedure, which averages across multiple rounds of validation taking as training and test sets different portions of the data (Ritchie, Kaplan, and Klein 2020, p. 587).

The outcome of the classification reveals how different populations of voxels encode different stimuli. This represents one of the main advantages of MVPA over univariate methods, given that the latter may fail to capture the processing of psychological variables that elicit multi-dimensional effects on neural activity. By analyzing the neural signal over a distributed pattern of voxels at the same time rather than in a single unit, MVPA results are more fine-grained than those from univariate methods and, thus, more sensitive in distinguishing the activity of those voxels that, even within the same brain region, do not carry identical information and respond differently to the same experimental variable (Davis, LaRocque, et al. 2014). Therefore, MVPA enriches the available neural information about the cognitive function under investigation, providing qualitatively better results than univariate methods.

For the previously mentioned reasons, it has been suggested that the pattern-based form of RI is generally more reliable than location-based RI. It can be schematized as follows (Nathan and Del Pinal 2017, p. 9):

P1. Cognitive process Cog_1 is associated by a classifier with a class of multi-voxel pattern Act_1, while Cog_2 is associated by a classifier with a class of multi-voxel pattern Act_2.
P2. In a neuroimaging experiment, during task *T*, multi-voxel pattern Act_1 is obtained.
C. During task *T*, process Cog_1 (rather than Cog_2) is engaged.

While the advantages of pattern-based RI over location-based RI are easily appreciated, studying cognitive states at the level of populations of voxels instead of at the level of individual voxels faces its own set of challenges. According to Ritchie, Kaplan, and Klein (2020), for instance, multivariate methods suffer from the "decoder's dictum" (p. 582), which is the assumption that the decoder (i.e., the classifier) reflects how the brain actually processes information.[8] Indeed, the simple fact that information can be accurately decoded from patterns of neural activity by means of a classifier is not a solid basis for inferring that those neural patterns actually *represent* the decoded information. Furthermore, Weiskopf (2021) notices that pattern-based RI is unable to fulfill the main theoretical aim of RI, viz., explaining the functional role of brain regions. Given the highly prediction-oriented purposes of classification algorithms, pattern-based RIs

[7] Of course, exactly how the classification is performed depends on the specific algorithm that is used. For example, many MVPA studies use linear classifiers such as linear support vector machines (for an introduction to the different classifiers used to analyze fMRI data, see Kriegeskorte and Bandettini 2007; Mahmoudi et al. 2012; Davis and Poldrack 2013; Wright 2018).

[8] See also DeWit et al. (2016) for a similar distinction between neuroimaging data as interpreted by experimenters ("experimenter-as-receiver") and by the brain ("cortex-as-receiver").

would not investigate the relationships between cognitive functions and their neural realizers, but would only offer a window on the potential biomarkers of cognitive processes. These are some of the issues motivating the explicit skepticism of a number of authors (e.g., Ritchie, Kaplan, and Klein 2020) about RI as based on MVPA methods, which do not seem sufficient to escape the weaknesses of conventional inferences in neuroscience.

3.3 Reverse Inference based on Representational Similarity Analysis

As mentioned above, it seems fair to say that the mere application of multivariate methods to the analysis of fMRI data is not enough to solve the methodological doubts concerning RI. However, as suggested by Ritchie, Kaplan, and Klein (2020) and Weiskopf (2021), specific multivariate methods can be instrumental in developing cognitive hypotheses with greater explanatory power, especially when enriched by further types of evidence such as behavioral and computational data. A case in point is a multivariate method known as "Representational Similarity Analysis" (RSA, for short; Kriegeskorte and Kievit 2013), which also presents interesting connections with the theoretical and methodological considerations from the recent philosophical and psychological literature on IBE to be discussed in the next section.[9]

The basic idea of RSA is to model measurements by means of a space whose dimensions reflect different features of the stimuli presented during the experiment. A virtue of RSA, as compared to other MVPA methods, is that it allows one to focus on more detailed and informative relationships holding between the data represented within the feature space.

RSA is a three-step procedure. First, the patterns of brain activation associated with each of the *N* stimuli presented during the experiment are arranged in a matrix. Second, the various neural patterns in the matrix are pairwise compared and their dissimilarities computed. This leads to a representational dissimilarity matrix (RDM), which can be spatially represented using a dimension-reduction technique such as multi-dimensional scaling or non-negative matrix factorization (e.g., Borg and Groenen 2005). An individual RDM comparing neural signals allows one to assess how a property putatively shared by a set of experimental stimuli is reflected in the activity of populations of voxels.

Up until this point, RSA appears similar to other MVPA methods. Characteristically, however, RSA involves different RDMs that can be constructed starting from the specific type of data available for the same stimuli; moreover, it does not necessarily require a classification task. Suppose, for instance, that researchers have conducted a behavioral experiment in which they recorded a certain behavioral variable (e.g., latencies, skin conductance responses, or similarity judgments) on the same stimuli used during the fMRI acquisition. Then, as a second step, a behavior-based RDM can be constructed for these responses.

The final step of RSA involves comparing the two (or more) RDMs in order to establish second-order dissimilarities between neural and behavioral data (Kriegeskorte, Mur, and Bandettini 2008; Kriegeskorte and Kievit 2013). While the neural and the behavioral RDMs respectively analyze the dissimilarities between neural and behavioral data for the stimuli at issue, the comparison of the two RDMs leads to a third, second-order matrix showing the dissimilarities between the dissimilarity patterns of the neural and the behavioral RDMs.[10] Therefore, rather

[9] For an introduction to the main differences between MVPA methods, see Kriegeskorte (2011) and Yang, Fang, and Weng (2012).

[10] The key idea of RSA reflects the concept of second-order isomorphism as first studied in Shepard (1968) and Shepard and Chipman (1970). Second-order isomorphism establishes a particular type of similarity between different kinds of data, such as behavioral responses and patterns of neural activity, as elicited by the same group of

than a simple association between properties of experimental stimuli and the (brain) responses to them, RSA allows researchers to establish, for those properties, the correspondences between patterns of behavioral responses and patterns of neural responses, providing a more detailed and integrated perspective for investigating the different pieces of evidence at hand.

There are no specific restrictions on the type of evidence to use within each RDM, leading to a variety of potential comparisons for the same stimuli, ranging from neural and behavioral data to computational models, recordings based on neuroimaging techniques other than fMRI (e.g., electroencephalography, magnetoencephalography), and interspecies data (Kriegeskorte, Mur, Ruff, et al. 2008). Therefore, RSA represents a remarkable method for connecting evidence from different areas of neuroscientific research (Kriegeskorte, Mur, and Bandettini 2008) and for developing richer explanations about the processing of a certain class of stimuli. In particular, as compared to other multivariate methods, RSA does not simply predict which experimental condition is more likely associated with the brain activity taken as input for a classifier but provides a detailed, integrated, and similarity-based model of how stimuli are processed.

The recent neuroscientific literature has widely discussed the pros and cons of univariate, multivariate, and RSA methods (see, among others, Davis, LaRocque, et al. 2014; Kriegeskorte 2011; Kriegeskorte and Kievit 2013; Hebart and Baker 2018). Here, we focus on the differences in the explanatory quality that studies based on RSA as compared to other methods may offer. We argue that, given its specific characteristics, RSA-backed hypotheses are more likely to satisfy the application criteria for IBE as put forward in the recent literature. In Section 5, we support this claim using a case study from the fMRI literature that relies on RSA. But, first, we give an overview of the most relevant parts of the discussion about IBE.

4 Inference to the Best Explanation and Explanatory Quality

Broadly put, IBE is a mode of inference grounded in the idea that explanation is a guide to belief, in the sense that if a hypothesis explains the available evidence better than its competitors, that gives reason to believe that the hypothesis is true. Philosophers have long thought this idea to be a cornerstone of modern scientific methodology, which they also took to provide all the justification of IBE one could ask for. But in the 1970s, some philosophers started to voice concerns over IBE, and in the 1980s and 1990s, especially with the advent of Bayesian confirmation theory, many philosophers of science had grown wary of IBE.

Presentations of IBE in older textbooks made it an easy target for critics. For instance, it was common to find IBE presented as licensing an inference to the truth of *that* member of a set of rival hypotheses that explains the available evidence best. But—critics (e.g., van Fraassen 1989) pointed out—what if the truth is not among the currently known candidate explanations? In general, there is no guarantee that we have been able to conceive of all possible explanations

experimental stimuli. Notably, RSA differs from methods relying on first-order isomorphism, allowing researchers to establish a relation between the property of a certain stimulus and the related brain or behavioral response (e.g., the eccentricity of an image in the visual field, its representation in the visual cortex, or its categorization during a behavioral task). Indeed, second-order isomorphism and, consequently, RSA allow researchers to analyze similarities between relations occurring among data of one type and relations occurring among data of another type, for the same set of stimuli (Kriegeskorte, Mur, and Bandettini 2008). A clear illustration of second-order isomorphism is the relationship occurring between a group of images ordered according to their eccentricities and their corresponding retinotopic representations in the visual cortex. While a comprehensive defense of RSA is beyond the scope of this work, it seems fair to say that RSA presents significant advantages over alternative methods for investigating multivariate data and supporting neuroscientific explanations.

of our evidence. To the contrary, often it will be reasonable to think we have *not*, as for instance argued by Stanford (2006) in his work on unconceived alternatives. Or what if the best available explanation strikes us as being still a quite unsatisfactory explanation of the evidence, or if perhaps it is fine, but there is another possible explanation that strikes us as being *almost* as fine? Should we still be happy, in those cases, to infer to the best explanation?

These and related, seemingly equally valid, concerns about IBE led its advocates to propose a number of fixes. Lipton (1993) argued that IBE should be understood as requiring that the best explanation be *good enough* for an inference to be warranted. And Bird (2010) added to this the requirement that the best explanation be *considerably better* than the second-best explanation.[11]

Of course, even if our best explanation appears excellent, and is also much more satisfactory than our second-best explanation, that in general is still no guarantee that the truth is among the hypotheses we are considering. But note that an inference to the best explanation need not be an all-or-nothing matter, in that we can make our confidence in that explanation proportional to our confidence that the truth is included in the designated set of hypotheses.

While plausible, this response creates problems of its own. For if IBE is conceived as a rule for determining degrees of confidence, it is in direct competition with Bayes' rule, the centerpiece of what has in the past decades become the dominant confirmation theory. According to this rule, upon the receipt of a piece of evidence, what until now were our degrees of confidence *conditional on* that piece of evidence occurring should become our new *un*conditional degrees of confidence. Bayesians have argued that *any* rule for changing one's degrees of confidence that is at variance with Bayes' rule is bound to lead to irrationality. Specifically, they have argued that anyone who changes their degrees of confidence via some non-Bayesian rule is liable to a dynamic Dutch book, that is, a set of bets, offered at various points in time, that all seem fair at the time they are offered but that collectively ensure a negative net pay-off. Using Bayes' rule instead—the rest of the argument goes—protects one from such bets. Given that—Bayesians claim—this liability can be figured out *a priori*, it is irrational to use any rule for changing degrees of confidence other than Bayes' rule (e.g., van Fraassen 1989). Another, currently more popular critique of non-Bayesian rules is that using such rules leads one to have degrees of confidence that should be expected to be less accurate than they could have been had one used Bayes' rule instead (with accuracy defined in terms of some so-called scoring rule; e.g., Greaves and Wallace 2006).

Both arguments have been contested, however, on two grounds. One is that they make unwarranted assumptions, or at least assumptions that opponents of Bayesianism need not buy into. For instance, the Dutch book argument assumes the Bayesian principle of expected utility maximization, which is controversial (see, e.g., Simon 1982), and the accuracy argument only takes one kind of accuracy (expected next-step accuracy) into account and fails if other, arguably more relevant, notions of accuracy are assumed (Douven 2022, Ch. 4). The second ground concerns a more elementary point, viz., that even granting both arguments, they only show that there are costs attached to using a non-Bayesian rule for changing one's degrees of confidence and entirely leave open the possibility that using such a rule has advantages that well outweigh the costs. And it has recently been shown that, in certain contexts, versions of IBE are indeed able to strike a better balance between two desiderata that typically pull in different

[11] In experimental studies, Douven and Mirabile (2018) found confirmation for the descriptive adequacy of Bird's and Lipton's amendments to IBE in that their participants were, *ceteris paribus*, reliably the more inclined to infer to the best explanation the better that explanation was, as judged by the participants, and were also, *ceteris paribus*, reliably the more inclined to infer to the best explanation the greater the difference in explanatory power between the best and the second-best explanation, again as judged by the participants.

directions, to wit, that of converging to the truth quickly (i.e., coming to hold a high degree of confidence in the truth quickly) and that of being accurate.[12]

It thus appears that we *can* consistently maintain that our confidence in the best explanation should reflect our confidence that the truth is among the candidate explanations we were able to conceive of. Arguably, that should not be the only determinant of our confidence in the best explanation. Once we think of IBE as a rule for changing degrees of confidence, rather than just asking whether the best explanation is good enough, and whether it is sufficiently much better than the second-best explanation, we should ask *how good* the best explanation is, the answer allowing for less and more, and similarly we should ask *how much better* than the second-best explanation the best explanation is.

Asking these questions makes a lot of sense, in fact, given how, according to philosophers of science, we are to judge explanatory goodness. For the properties that are supposed to factor in such judgments are the so-called theoretical virtues, which are almost all graded, and not categorical. To be sure, we want an explanation to be consistent with the evidence, and consistency is, on all of the better-known logics, an all-or-nothing matter. But explanatory goodness is also commonly assumed to be a matter of coherence, most notably coherence with background knowledge, and coherence famously permits of degrees (Bovens and Hartmann 2003; Douven and Meijs 2007).

A theoretical virtue that merits especial emphasis, because it will be particularly relevant in the next section, is what Whewell (1847) calls "consilience of inductions," by which he means the finding that a theory is able to explain several bodies of data that previously appeared unconnected to each other. If it happens, that brings about a "unification of two or more hitherto disparate areas of understanding beneath one or a few high-level hypotheses or established laws" (Ruse 1975, p. 2 f), which can be taken as an indication that those hypotheses or laws are true. As Ruse (1975) and Thagard (1977) note, this idea had a strong influence on Darwin, who argued for his theory of individual variation and selective retention precisely on the basis that the theory was able to explain a great variety of what had appeared to be completely disparate facts (Darwin 1876, p. 421). Just like coherence and simplicity can be realized to different degrees, consilience of inductions can, because the number of different areas that the explanation is able to connect can vary, and also because these areas can have appeared independent from one another to different degrees before we came to see them as connected, due to the new explanation.

What all of this shows, we believe, is that IBE is best thought of *not* as a rule that invariably compels us to infer the truth of the best explanation of our evidence. Despite its name, IBE need not involve an outright inference but may instead make us more confident in the best explanation, where the exact degree of confidence can depend on many factors: how coherent the best explanation is with background knowledge, how simple it is, how varied the evidence is it is able to explain, how much better it is than the next best alternative, and more.

From this, it should also be clear that, whichever epistemic attitude an application of IBE warrants, that attitude is revisable, again for a number of reasons. Most obviously, we may be able to come up with a better explanation still, and even if not, we may be able to come up with a rival that is about as good, *qua* explanation, which may weaken our confidence in the best explanation.

[12] Douven (2022, Chs. 6 and 7) shows that whether or not a probabilistic version of IBE beats Bayes' rule in the said respect depends on how it assigns bonus weights for explanatory quality; the details of these versions need not detain us here, however.

We mentioned previously that we view RI as a form of IBE, and we believe all of the foregoing to apply to RI as well. Specifically, we believe that it is wrong to see RI as necessarily involving an outright inference. Rather, it is a principle that helps us determine how much confidence to invest in a cognitive hypothesis given our evidence, and taking into account at least two key factors, namely,

1. how well the hypothesis explains the evidence, specifically, whether it is a sufficiently good explanation; and
2. how much better (if at all) the hypothesis explains the evidence than its rivals, specifically, for a best-explaining hypothesis, whether it is sufficiently much better than the second-best explanation.

And in answering these two questions, we will especially want to consider how well the given hypothesis coheres with existing knowledge and also how varied the body of evidence is that it is able to explain. To make the proposed view on RI concrete, we look in some detail at two specific applications of this form of inference.

5 Reverse Inference as Inference to the Best Explanation: Two Case Studies

In the previous section, we summarized recent philosophical work on abduction to isolate the factors that contribute to the credibility of the hypothesis that best explains the available evidence. These factors are conceived as general virtues meant to be compatible with any model of explanation—whether mechanistic, causal, unificationist, teleological, or otherwise—discussed by philosophers of science, philosophers of mind and cognitive neuroscientists. In line with this, the aim of our analysis is to scrutinize the degree to which adhering to certain principles, such as Whewell's consilience of inductions, can assure the quality of a given explanation, regardless of the favored model of explanation.[13]

In the present section, we discuss and compare two different case studies from recent neuroscientific research. While both aim at supporting, through an RI, a specific hypothesis concerning the involvement of a cognitive process as an explanation of the available evidence, the degree of confidence in their conclusions differs widely, precisely because the principles highlighted in the previous section are satisfied either not at all or to a much lesser extent by one study rather than the other.

5.1 Kriegeskorte et al.'s studies on object categorization

In two separate studies, Kriegeskorte and colleagues investigated the role of the inferior temporal cortex (ITC) in object representation and categorization by comparing neuroimaging data from macaque monkeys with neuroimaging data from humans (Kriegeskorte, Mur, Ruff, et al. 2008; Kriegeskorte, Mur, and Bandettini 2008). Relying on previous research, the authors hypothesized that this region plays an analogous role in humans and monkeys and, in particular, that it is possible to infer the presence of a process of object categorization from ITC activity.

[13] See Douven (2022, Ch. 1) and Prasetya (forthcoming) for arguments to the effect that proponents of IBE, as understood here (i.e., as serving ampliative purposes), need not commit to any specific model of explanation. See also Colombo (2017) for why we should be open to pluralism with regard to models of explanation. Therefore, our proposal is compatible with various accounts of explanation that weigh explanatory virtues differently. However, the detailed discussion of this aspect extends beyond the scope of the present work.

To test their hypothesis, these authors made crucial use of RSA. They designed and ran a neuroimaging experiment in which single-cell recording and fMRI are used to register the neural activity, respectively, of monkeys and humans during the presentation of a set of pictures representing isolated real-world objects, including natural and artificial inanimate objects and faces or body parts of humans and nonhuman animals (Kriegeskorte, Mur, Ruff, et al. 2008, p. 1138). Researchers utilized single-cell recording to measure changes in voltage or current extracellularly, through an electrode implanted into the animal's skull close to the area of interest. The recorded signals of each stimulus were pairwise compared and an RDM showing the dissimilarity of the single-cell recordings elicited in the ITC of monkeys was calculated. The very same strategy was used to analyze fMRI-based data from the human ITC. Then another, independent, human-related RDM was built.

The RSA based on the resulting RDMs showed a matching between neural data from monkeys and from humans, in other words, the dissimilarities among the representations of experimental stimuli in monkey ITC reflected analogous trends in human ITC. As attested by the authors, results this clear were unexpected both for the different types of techniques used to acquire data in monkeys and humans and for the original, inter-species method employed (p. 1128). However, these inter-species results appear to strongly support Kriegeskorte and colleagues' explanation of the role of the ITC in processing object categories across species.

A look at Kriegeskorte, Mur, and Bandettini (2008) brings the potential of RSA further into relief. In this work, the authors took the set of stimuli for the experiment reported in Kriegeskorte, Mur, Ruff, et al. (2008) and used them as input for several computational models. The aim of this new analysis was to explore how similarity patterns found at the level of neural regions both in humans and monkeys are reflected in the processing steps of artificial models meant to simulate parts of the visual process. For instance, the authors used a model of the primary visual cortex employing a series of Gabor filters[14] for analyzing spatial frequencies and orientations of stimuli (Kriegeskorte, Mur, and Bandettini 2008, p. 7) as well as other models implementing more complex categorical discrimination between pictures, such as the "animate–inanimate model," which classifies two stimuli as being alike if they are either both animate or both inanimate and else as different (Kriegeskorte, Mur, and Bandettini 2008, p. 7). From each model, an RDM was obtained and further compared to neural and behavioral RDMs obtained from the same experimental stimuli to detect potential similarity patterns. The comparative analysis of the various RDMs reveals the importance of RSA in underlining potential similarities between artificial and biological processing of the stimuli, for instance, by detecting whether information representations in specific models, or parts of them (e.g., specific layers in a Deep Neural Network), resemble those in human brain regions or particular portions of them (Kriegeskorte, Mur, and Bandettini 2008, p. 16).

These two studies by Kriegeskorte and colleagues offer a particularly interesting case for discussing some aspects from the recent literature on IBE and the notion of explanation. Indeed, we believe the aforementioned studies do exceptionally well in light of the considerations discussed in Section 4, and to do much better than other studies in the same field. To support our claim, we below briefly discuss in turn the criteria of explanatory quality mentioned in Section 4, emphasizing how they are crucially met in the studies of Kriegeskorte and colleagues as compared to other studies from the same field.

[14] A Gabor filter is a specific transformation used in image processing for texture analysis and feature extraction. Many studies in visual neuroscience (e.g., Kay et al. 2008) use Gabor filters for modeling the activity of voxels within the primary visual cortex associated with the processing of visual features.

First, Kriegeskorte and colleagues' hypothesis about the interspecies role of the ITC in visual processing and category discrimination coheres well with background knowledge accumulated since the early stages of neuroimaging research (see, among others, Puce et al. 1995; Kanwisher, McDermott, and Chun 1997; Haxby, Gobbini, et al. 2001; Van Essen et al. 2001; Tsao et al. 2003; Kiani et al. 2007) and has been further confirmed by recent studies (e.g., Huth et al. 2012; Bao et al. 2020). That research found evidence supporting the hypothesis that both human ITC and monkey ITC are sensitive to the processing of real-world objects, by associating the average neural response registered at the level of the ITC to stimuli of predefined object categories, such as human faces. These results motivated the specific question addressed by Kriegeskorte, Mur, Ruff, et al. (2008), that is, whether the ITC—both in humans and in monkeys—does not simply respond to specific real-world object categories but plays a more general function in processing categorical knowledge, rather than purely visual properties about objects. Indeed, Kriegeskorte and colleagues' conclusion that the dominant factor determining ITC activity is the category of the stimulus and, more broadly, semantic and higher-level properties of objects (such as their animacy; Kriegeskorte, Mur, Ruff, et al. 2008, p. 1127 f) seems to provide a quite detailed and complete understanding of the role of the ITC, given the available evidence and the ongoing research.

However, as noticed by Conway (2018, p. 12), the extent to which semantic and higher-level properties of objects (e.g., animacy) should be considered as the organizing principle of the activity of the ITC is not completely uncontroversial in the literature. Indeed, Baldassi et al. (2013) argue that the ITC processes reflect similarity rather than categorical membership of objects, that is, the activity of the ITC is better accounted for by lower-level visual properties of stimuli, such as their shape and the presence of specific geometrical patterns. While the conclusion of Kriegeskorte and colleagues does not seem to encounter any rival explanation regarding the *general* role of the ITC in object discrimination, once the specific type of information the ITC is supposed to process comes into question, other explanations do become available, the hypothesis provided by Baldassi et al. (2013) being a particularly notable rival to Kriegeskorte and colleagues' proposal.

The two studies are also very similar from a methodological perspective. Both refer to very similar datasets of experimental stimuli (i.e., pictures of isolated real-world objects from different categories), rely on electrophysiological recordings from monkeys as evidence, and use multivariate approaches to analyze their data (RSA in the case of Kriegeskorte and colleagues and a machine-learning approach based on an unsupervised clustering algorithm in Baldassi and colleagues). And as Conway (2018) shows, the work reported in Baldassi et al. (2013) also coheres well with background knowledge. So far, we would seem to have two good yet conflicting explanations of specific activities of ITC.

However, a key difference, which makes Kriegeskorte et al.'s explanation a clear winner in our eyes, concerns the variety of evidence the two studies take into account. While Baldassi and colleagues analyze neuronal recordings from monkeys only, Kriegeskorte and colleagues are able to merge different types of data, such as animal-based neuronal recordings, human fMRI, and data from computational models, through RSA. It is worth elaborating on this virtue of Kriegeskorte et al.'s work.

The authors infer from the close match between human and primate RDMs that neural patterns within the ITC respond to conventional category boundaries (e.g., animate versus inanimate objects) and process features that appear interspecifically relevant for grouping stimuli into different categories. In particular, the ITC activity reveals—both in humans and

monkeys—a categorical distinction between animate-related and inanimate-related stimuli, and, among the former, between stimuli depicting faces and those depicting body parts. The importance of ITC for categorical representation is further highlighted by evidence from artificial models meant to simulate parts of the visual process (Kriegeskorte, Mur, and Bandettini 2008). Most notably, a comparison between the RDMs of computational models and the neural RDMs associated with brain regions involved in different steps of visual processing—that is, early visual cortices located outside ITC and the fusiform face area (FFA), within the ITC—shows that complex artificial models proposed for simulating category discrimination match the response of the FFA (and thus of the ITC) better than simpler models used for processing low-level visual information (whose RDMs, on the contrary, fit better the RDMs based on the activity of early visual brain regions, located outside the ITC).

The strength of Kriegeskorte and coauthors' work is due, to a large extent, to the fact that they use the RSA method in their data analysis, specifically, that they use it as a crucial tool for connecting evidence from different neuroscientific sub-fields. As emphasized in Section 3.3, while other inferential patterns in cognitive neuroscience only rely on neural evidence, RSA is explicitly meant to take into account different types of evidence, specifically, data that are acquired via different techniques and that, independently from each other, may point to the same explanation. Even though fMRI data are radically different from neuronal recordings (Kriegeskorte, Mur, Ruff, et al. 2008, p. 1128) and outcomes of computational models, provided the experimental stimuli are the same across techniques and models, RSA offers a framework for assessing the convergence of distinct pathways of reasoning based on different types of evidence. In other words, RSA allows researchers to determine whether seemingly independent pieces of evidence support the very same explanation, typically relying on different kinds of observation at the cognitive, behavioral, and computational level. Thereby, RSA can establish the kind of consilience of inductions that Whewell saw as indicating the truth of a hypothesis.[15]

As seen in Section 4, a relevant caveat for inferring to the best explanation is represented by the impact of competing hypotheses, and how much more satisfactory the target explanation appears than its competitors. As for our case study, the conclusion of Baldassi et al. (2013) appears weaker, in terms of explanation quality, than the conclusion provided by Kriegeskorte and colleagues. To assess whether object representation in the ITC depends either on the semantic membership of objects, their shape features or other, low-level visual properties, Baldassi et al. (2013, p. 18) rely only on the outcome of different unsupervised clustering algorithms trained on neural recordings from monkeys. When Baldassi et al. (2013) and Kriegeskorte and colleagues' 2008 study are compared in terms of richness and variety of evidence considered, it is clear that the explanation provided by the latter does substantially better than that provided by the former.

To be clear, that Kriegeskorte et al.'s hypothesis is supported via a successful application of RSA does not exclude the possibility that hypotheses will emerge which are able to explain the current relevant data better still. Also, we may obtain further data which will favor other hypotheses over Kriegeskorte et al.'s. As a result, we could revise our verdict about Kriegeskorte et al.'s conclusion, or could at least lose some of our confidence in it. But this is no different than for any other form of non-deductive inference.

[15] RSA is not the only method that matches the criteria we discussed in Section 4, nor is it necessarily the best; some other notable multi-method approaches are mentioned in Section 5.2. Nevertheless, we do believe that, when compared to these other approaches, RSA represents the clearest illustration of how the criteria contributing to the quality of IBEs are applied in neuroscience.

5.2 Lieberman and Eisenberger's study on pain

To better illustrate the extent to which, in line with viewing RIs as IBEs, the confidence to invest in a cognitive hypothesis depends on the factors presented in Section 4, it is useful to discuss another study that recently triggered a critical debate among neuroscientists.

In a paper titled "The dorsal anterior cingulate cortex is selective for pain: Results from large-scale reverse inference," Lieberman and Eisenberger (2015) claim to have shown that the dorsal anterior cingulate cortex (dACC), a region generally associated with a variety of cognitive functions, is actually mainly devoted to pain processing. Interestingly, the authors do not perform experiments, but run a meta-analysis using the tools provided by NeuroSynth (henceforth NS), a large-scale automated platform that synthesizes data from more than 14,000 published fMRI studies (Yarkoni et al. 2011).

In a nutshell, NS works as follows. Using data-mining algorithms applied to the papers in its database, NS automatically extracts relevant cognitive terms (such as "language" or "working memory") appearing in those papers together with the coordinates corresponding to brain activations as typically reported in the papers' figures. The former are used as proxies of the cognitive functions and processes supposedly investigated in the papers; the latter, as possible loci of neural realization of those processes. Then, NS provides summary statistics (z-scores and posterior probabilities) of the associations between brain activations and terms over the database. More specifically, it implements the Bayesian analysis of RI proposed by Poldrack (2006, 2011) in order to estimate the probability $\Pr(Cog \mid Act)$ that a term *Cog* (as proxy of a certain cognitive function) occurs in papers reporting the brain activation *Act*, as based on the likelihood $\Pr(Act \mid Cog)$ and assuming an uninformative prior (i.e., that a priori *Cog*'s occurring is as probable as not). NS has been specifically designed—partly in response to the worries raised by Poldrack (2006) and others—to support large and systematic RIs, and it is now widely employed by the community to design experiments and evaluate hypotheses about cognitive functions (for a survey, see Coraci, Calzavarini, and Cevolani 2022).

In their study, Lieberman and Eisenberger recruit NS to show that, for papers reporting the activation of voxels within the dACC, pain-related terms had a higher probability of being present in the text of papers as compared to other terms, suggesting that the psychological state that can be reliably inferred from dACC activity is pain (Lieberman and Eisenberger 2015, p. 15252). This claim attracted much criticism, with several scholars (see, e.g., Yarkoni 2015a,b; Wager et al. 2016) arguing that such a conclusion was probably mistaken and in any case not adequately justified (for a review of the debate, see Calzavarini and Cevolani 2022). Without going into the details of this discussion, we note that, by design, a meta-analysis combines results from multiple studies that focus on the same research question, in order to spot potential convergence or disagreement between them. Therefore, despite the criticisms that have been leveled at this methodology (e.g., Stegenga 2011), meta-analyses of the literature represent one of the main strategies available to researchers to achieve a form of consilience of inductions, that is, gathering different, independent experiments and provide convergent evidence toward the same explanation. Still, as testified by the strongly critical reactions mentioned above, Lieberman and Eisenberger's analysis failed to gain the approval of the community. It also presents an interesting case to further discuss the determinants of the quality of an IBE discussed in Section 4.

Grant that Lieberman and Eisenberger's hypothesis—that the dACC is selective for pain—is coherent with background knowledge about the cognitive functions engaged during dACC activity, and also that it can be considered a sufficiently good explanation of the available

evidence, given both the number of pain-related studies from the NS database that support such a conclusion and the significance of the statistical tests the authors conducted.[16] Then their proposal still fails to satisfy the criteria of explanation quality that, as argued in the previous section, *are* met by Kriegeskorte et al.'s analysis.

In particular, and as pointed out by several authors, while the dACC is surely *associated* with pain, to conclude that the dACC is *selective* for pain—that is, that the best explanation for the patterns we see in the data is that the function of the dACC is processing of pain stimuli and/or generating pain responses—is highly questionable, given that many fMRI studies appear to show that data indicating dACC activity is also compatible with other candidate explanations, such as associations to empathy, anxiety disorders, abuse, and dysregulation (see Yarkoni 2015a,b; Wager et al. 2016). In view of this, the best explanation may even be a "disjunctive" one, pointing to two or more cognitive processes and outcompeting any single "monofunctional" hypothesis. Thus, differently from the case of Kriegeskorte and colleagues' hypothesis about the activity of the ITC, Lieberman and Eisenberger's pain-related hypothesis—as an explanation for dACC activation—does *not* seem to be by far better than all other candidates. Also, Lieberman and Eisenberger cannot appeal to a variety of evidence their hypothesis would be able to explain either, for even though the meta-analysis that these authors ran by means of NS takes into account many independent experiments, it still considers only fMRI data. As a result, their analysis does not bring about the kind of consilience of inductions found in Kriegeskorte et al.'s work which, relying on the RSA methodology, exploits the coherence of neural, inter-species, behavioral, and computational evidence and thereby shows clear explanatory advantages when compared to studies proposing rival hypotheses.

In closing, it is worth mentioning more recent work from the same lab, in which Lieberman and colleagues (Lieberman, Straccia, et al. 2019) report an analysis similar to the one of Lieberman and Eisenberger. In the new analysis, these authors draw an RI about the functional role of the medial prefrontal cortex in social cognition and, in particular, social, self-related, and affective processes. However, in contrast with Lieberman and Eisenberger (2015), the authors followed a multi-method approach to support their conclusion. In addition to using meta-analytic evidence from NS, they assess the reliability of the associations between the medial prefrontal cortex and the cognitive domains of interest by reviewing evidence from different studies employing distinct methods: lesion works, studies based on transcranial magnetic stimulation techniques, and fMRI studies using multivariate pattern analyses. From our current perspective, this strategy of supporting neuroscientific hypotheses and strengthening the reliability of neuroscientific inferences by relying on various and integrated types of evidence can be interpreted as aiming at a consilience of inductions, discussed above as a theoretical virtue central to assessing how good an explanation is. Thus, seeing the criteria for successful applications of IBEs as applying to RIs as well, we would argue that the use of RI is much more warranted in Lieberman, Straccia, et al. (2019) than it is in Lieberman and Eisenberger (2015).

6 Conclusion

While commonly used in neuroscience, the methodological status of RI has been called into question, with some even likening this type of inference to a logical fallacy. We have proposed a different interpretation of RI, one on which it is an instance of a graded form of IBE, according

[16] It is to be noted, though, that these tests have been extensively criticized; see, e.g., Yarkoni (2015a,b) and Wager et al. (2016).

to which we are licensed to have confidence in the truth of the cognitive hypothesis that explains the neural, and possibly other, evidence best. Taking inspiration from recent work on IBE, we argued that the *degree* of our confidence in that hypothesis should depend on a variety of factors, such as how satisfactory the hypothesis is *qua* explanation, how much better it is, *qua* explanation, than its closest competitors, how confident we are that we are not overlooking some superior explanation, and so on.

This means that concrete RIs to be found in the literature should be neither rejected a priori nor endorsed a priori by defending them as IBEs *simpliciter*. Instead, how compelling each of them is qua IBE, is to be determined on a case-by-case basis, in light of the criteria discussed in Section 4, and taking into account the characteristics of the relevant experimental data and the specific methodology used for analyzing them. Thus, RI is more plausibly viewed as a context-dependent principle than as a universal principle that applies across the board. To buttress our proposal, we canvassed Kriegeskorte and colleagues' studies, whose use of RI did well in light of all relevant criteria for IBE, and we contrasted those studies with another (by Lieberman and Eisenberger) in which—we argued—the use of RI should lead to a more guarded conclusion.

Acknowledgments. We current work has been presented at the Center for Logic, Language, and Cognition (University of Turin) and during the 2022 Annual Conference of the Italian Association of Cognitive Science. We are greatly indebted to Christopher von Bülow for valuable comments on a previous version. We would also like to thank Fabrizio Calzavarini, Luca Cecchetti, Vincenzo Crupi, and Jan Sprenger for very useful discussions on the topics of the chapter. Davide Coraci acknowledges funding from an ERASMUS+ Mobility Grant 2020/2021; Gustavo Cevolani from the Italian Ministry of University and Research (MUR) through the PRIN 2022 grant n. 2022ARRY9N (Reasoning with hypotheses: Integrating logical, probabilistic, and experimental perspectives), funded by the European Union (Next Generation EU).

References

Anderson, M. L. (2010). "Review of *Neuroeconomics: Decision making and the brain*." In: *Journal of Economic Psychology* 31, pp. 151–154.

Baldassi, C. et al. (2013). "Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons." In: *PLoS Computational Biology* 9.8, e1003167.

Bao, P. et al. (2020). "A map of object space in primate inferotemporal cortex." In: *Nature* 583.7814, pp. 103–108.

Bird, A. (2010). "Eliminative abduction: Examples from medicine." In: *Studies in the History and Philosophy of Science* 41, pp. 345–352.

Borg, I. and P. J. F. Groenen (2005). *Modern multidimensional scaling*. 2nd edition. New York: Springer.

Bourgeois-Gironde, S. (2010). "Is neuroeconomics doomed by the reverse inference fallacy?" In: *Mind & Society* 9, pp. 229–249.

Bovens, L. and S. Hartmann (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Boyd, R. N. (1984). "On the current status of scientific realism." In: *Erkenntnis* 19, pp. 45–90.

Burnston, D. C. (2016). "A contextualist approach to functional localization in the brain." In: *Biology & Philosophy* 31, pp. 527–550.

Calzavarini, F. and G. Cevolani (2022). "Abductive reasoning in cognitive neuroscience: Weak and strong reverse inference." In: *Synthese* 200.2, pp. 1–26.
Colombo, M. (2017). "Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation." In: *Cognitive Science* 41.2, pp. 503–517. DOI: 10.1111/cogs.12340.
Conway, B. R. (2018). "The organization and operation of inferior temporal cortex." In: *Annual Review of Vision Science* 4, pp. 381–402.
Coraci, D., F. Calzavarini, and G. Cevolani (2022). "Reverse inference, abduction, and probability in cognitive neuroscience." In: *Handbook of abductive cognition*. Ed. by L. Magnani. Cham: Springer, pp. 1–27. DOI: 10.1007/978-3-030-68436-5_60-1.
Costa, T. et al. (2021). "BACON: A tool for reverse inference in brain activation and alteration." In: *Human Brain Mapping* 42.11, p. 3343.
Darwin, C. (1876). *On the origin of species by means of natural selection*. 6th edition. London: John Murray.
Davis, T., K. F. LaRocque, et al. (2014). "What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis." In: *Neuroimage* 97, pp. 271–283.
Davis, T. and R. A. Poldrack (2013). "Measuring neural representations with fMRI: Practices and pitfalls." In: *Annals of the New York Academy of Sciences* 1296.1, pp. 108–134.
Deeley, J. (1994). *The collected papers of Charles Sanders Peirce*. URL: https://colorysemiotica.files.wordpress.com/2014/08/peirce-collectedpapers.pdf.
Del Pinal, G. and M. J. Nathan (2017). "Two kinds of reverse inference in cognitive neuroscience." In: *The human sciences after the decade of the brain*. Ed. by J. Leefmann and E. Hildt. Oxford Academic Press, pp. 121–139.
DeWit, L. et al. (2016). "Is neuroimaging measuring information in the brain?" In: *Psychonomic Bulletin & Review* 23, pp. 1415–1428.
Douven, I. (2022). *The art of abduction*. Cambridge MA: MIT Press.
Douven, I. and W. Meijs (2007). "Measuring coherence." In: *Synthese* 156, pp. 405–425.
Douven, I. and P. Mirabile (2018). "Best, second-best, and good-enough explanations: How they matter to reasoning." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44.11, pp. 1792–1813.
Glymour, C. and C. Hanson (2016). "Reverse inference in neuropsychology." In: *The British Journal for the Philosophy of Science* 67.4, pp. 1139–1153.
Greaves, H. and D. Wallace (2006). "Justifying conditionalization: Conditionalization maximizes expected epistemic utility." In: *Mind* 115, pp. 607–632.
Haxby, J. V., A. C. Connolly, and J. S. Guntupalli (2014). "Decoding neural representational spaces using multivariate pattern analysis." In: *Annual Review of Neuroscience* 37.1, pp. 435–456.
Haxby, J. V., M. I. Gobbini, et al. (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex." In: *Science* 293.5539, pp. 2425–2430.
Haynes, J. D. and G. Rees (2006). "Decoding mental states from brain activity in humans." In: *Nature Reviews Neuroscience* 7.7, pp. 411–421.
Hebart, M. N. and C. I. Baker (2018). "Deconstructing multivariate decoding for the study of brain function." In: *Neuroimage* 180, pp. 4–18.
Huth, A. G. et al. (2012). "A continuous semantic space describes the representation of thousands of object and action categories across the human brain." In: *Neuron* 76.6, pp. 1210–1224.
Hutzler, F. (2014). "Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data." In: *Neuroimage* 84, pp. 1061–1069.

Kanwisher, N., J. McDermott, and M. M. Chun (1997). "The fusiform face area: A module in human extrastriate cortex specialized for face perception." In: *Journal of Neuroscience* 17.11, pp. 4302–4311.

Kay, K. N. et al. (2008). "Identifying natural images from human brain activity." In: *Nature* 452.7185, pp. 352–355.

Kiani, R. et al. (2007). "Object category structure in response patterns of neuronal population in monkey inferior temporal cortex." In: *Journal of Neurophysiology* 97.6, pp. 4296–4309.

Kriegeskorte, N. (2008). "Representational similarity analysis –- connecting the branches of systems neuroscience." en. In: *Frontiers in Systems Neuroscience*. ISSN: 1662-5137. DOI: 10.3389/neuro.06.004.2008. (Visited on 01/22/2020).

— (2011). "Pattern-information analysis: From stimulus decoding to computational-model testing." In: *Neuroimage* 56.2, pp. 411–421.

Kriegeskorte, N. and P. A. Bandettini (2007). "Analyzing for information, not activation, to exploit high-resolution fMRI." In: *Neuroimage* 38.4, pp. 649–662.

Kriegeskorte, N. and R. A. Kievit (2013). "Representational geometry: Integrating cognition, computation, and the brain." In: *Trends in Cognitive Sciences* 17.8, pp. 401–412.

Kriegeskorte, N., M. Mur, and P. A. Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience." In: *Frontiers in Systems Neuroscience*, pp. 1–28.

Kriegeskorte, N., M. Mur, D. A. Ruff, et al. (2008). "Matching categorical object representations in inferior temporal cortex of man and monkey." In: *Neuron* 60.6, pp. 1126–1141.

Lieberman, M. D. and N. I. Eisenberger (2015). "The dorsal anterior cingulate cortex is selective for pain: Results from large-scale reverse inference." In: *Proceedings of the National Academy of Sciences* 112.49, pp. 15250–15255.

Lieberman, M. D., M. A. Straccia, et al. (2019). "Social, self (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence." In: *Neuroscience & Biobehavioral Reviews* 99, pp. 311–328.

Lipton, P. (1993). "Is the best good enough?" In: *Proceedings of the Aristotelian Society* 93, pp. 89–104.

Machery, E. (2014). "In defense of reverse inference." In: *The British Journal for the Philosophy of Science* 65.2, pp. 251–267.

Mahmoudi, A. et al. (2012). "Multivoxel pattern analysis for fMRI data: A review." In: *Computational and Mathematical Methods in Medicine* 2012.

McMullin, E. (1992). *The inference that makes science*. Milwaukee WI: Marquette University Press.

Nathan, M. J. and G. Del Pinal (2017). "The future of cognitive neuroscience? Reverse inference in focus." In: *Philosophy Compass* 12.7, e12427.

Norman, K. A. et al. (2006). "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data." In: *Trends in Cognitive Sciences* 10, pp. 424–430.

Poldrack, R. A. (2006). "Can cognitive processes be inferred from neuroimaging data?" In: *Trends in Cognitive Sciences* 10.2, pp. 59–63.

— (2011). "Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding." In: *Neuron* 72.5, pp. 692–697.

Prasetya, Y. (forthcoming). "Which models of scientific explanation are (in)compatible with IBE?" In: *British Journal for the Philosophy of Science*.

Puce, A. et al. (1995). "Face-sensitive regions in human extrastriate cortex studied by functional MRI." In: *Journal of Neurophysiology* 74.3, pp. 1192–1199.

Ritchie, J. B., D. M. Kaplan, and C. Klein (2020). "Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience." In: *The British Journal for the Philosophy of Science* 70.2, pp. 581–607.
Ruse, M. (1975). "Darwin's debt to philosophy: An examination of the influence of the philosophical ideas of John F. W. Herschel and William Whewell on the development of Charles Darwin's theory of evolution." In: *Studies in the History and Philosophy of Science* 6, pp. 159–181.
Shepard, R. N. (1968). "Book review: *Cognitive psychology* by Ulric Neisser." In: *The American Journal of Psychology* 81.2, pp. 285–289.
Shepard, R. N. and S. Chipman (1970). "Second-order isomorphism of internal representations: Shapes of states." In: *Cognitive Psychology* 1.1, pp. 1–17.
Simon, H. A. (1982). *Models of bounded rationality, Vol. I*. Cambridge MA: MIT Press.
Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford: Oxford University Press.
Stegenga, J. (2011). "Is meta-analysis the platinum standard of evidence?" In: *Studies in History and Philosophy of Science Part C* 42.4, pp. 497–507.
Thagard, P. R. (1977). "Discussion: Darwin and Whewell." In: *Studies in the History and Philosophy of Science* 8, pp. 353–356.
Tsao, D. Y. et al. (2003). "Faces and objects in macaque cerebral cortex." In: *Nature Neuroscience* 6.9, pp. 989–995.
Van Essen, D. C. et al. (2001). "Mapping visual cortex in monkeys and humans using surface-based atlases." In: *Vision Research* 41.10-11, pp. 1359–1378.
van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.
Wager, T. D. et al. (2016). "Pain in the ACC?" In: *Proceedings of the National Academy of Sciences* 113.18, E2474–E2475.
Weiskopf, D. A. (2021). "Data mining the brain to decode the mind." In: *Neural mechanisms*. Ed. by F. Calzavarini and M. Viola. Cham: Springer, pp. 85–110.
Whewell, W. (1847). *The philosophy of the inductive sciences, founded upon their history*. London: John W. Parker.
Wright, J. (2018). "The analysis of data and the evidential scope of neuroimaging results." In: *The British Journal for the Philosophy of Science*.
Yang, Z., F. Fang, and X. Weng (2012). "Recent developments in multivariate pattern analysis for functional MRI." In: *Neuroscience Bulletin* 28.4, pp. 399–408.
Yarkoni, T. (2015a). *No, the dorsal anterior cingulate is not selective for pain: Comment on Lieberman and Eisenberger*. URL: https://www.%20talyarkoni.%20org/blog/2015a/12/05/no-the-dorsal-anterior-cingulate-is-not-selective-for-pain-comment-on-lieberman-and-eisenberger-2015a.
— (2015b). *Still not selective: Comment on comment on comment on Lieberman & Eisenberger*. URL: https://www.%20talyarkoni.%20org/blog/2015b/12/14/still-not-selective-comment-on-comment-on-comment-on-lieberman-eisenberger-2015b.
Yarkoni, T. et al. (2011). "Large-scale automated synthesis of human functional neuroimaging data." In: *Nature Methods* 8.8, pp. 665–670.