Realism, underdetermination, and inference in cognitive neuroscience

Davide Coraci, Gustavo Cevolani
IMT School for Advanced Studies Lucca
davide.coraci@imtlucca.it,gustavo.cevolani@imtlucca.it

Abstract. In this paper, the realism/antirealism debate is discussed within the context of cognitive neuroscience, with a particular focus on experimental work that utilizes functional magnetic resonance imaging (fMRI). We start with a general discussion of realism and antirealism in cognitive science. Then, we consider the recent debate surrounding reverse inference, a reasoning pattern utilized by neuroscientists to support hypotheses about the engagement of cognitive processes based on patterns of brain activity observed through fMRI. We argue that reverse inference poses a problem of empirical underdetermination with respect to the theoretical interpretation of experimental results in cognitive neuroscience. As a result, it seemingly presents an antirealist argument regarding the evaluation of cognitive hypotheses. However, a closer examination of recent literature reveals that neuroscientists are successfully addressing the problem of reverse inference in a variety of ways that are closely related to how a philosophical defense of realism in the face of underdetermination would operate. These findings provide evidence for a qualified, realist interpretation of current cognitive neuroscience methodology.

Keywords: scientific realism; empirical underdetermination; reverse inference; fMRI; theoretical virtues; Bayesianism; abduction; cognitive neuroscience.

The authors wish to thank Sara Dellantonio and Luca Tambolo for helpful discussion, and two anonymous reviewers for their comments on the manuscript. Financial support from the Italian Ministry of Scientific Research (PRO3 project "Understanding public data: experts, decisions, epistemic values", CUP D67G22000130001) is gratefully acknowledged.

1 Introduction

Realists assert that it is justifiable to commit to our best scientific hypotheses, and to believe in both observable and unobservable aspects of the world as described by these hypotheses. Antirealists, on the other hand, disagree with this optimistic view and put forth several arguments to undermine it.

The philosophical literature exhibits a full range of such arguments and counterarguments, with the problem of empirical underdetermination playing a central role among others. This paper discusses the realism/antirealism debate within the context of cognitive neuroscience, with a particular focus on experimental work utilizing functional magnetic resonance imaging (fMRI). Section 2 reviews some well-known debates in cognitive science to illustrate how a broadly realistic view arguably underlies much current research in this field. Section 3 focuses on the recent debate concerning the issue of "reverse inference," interpreting it as an instance of empirical underdetermination, possibly favoring an antirealist view of neuroscientific methodology. Against this impression, Section 4 demonstrates how both neuroscientific and philosophical discussion of reverse inference appear to favor a realistic interpretation of scientific practice in this field. Two examples of such discussion are analyzed, highlighting how they closely parallel traditional defenses of scientific realism. A brief, tentative conclusion is offered in Section 5.

2 Realism and antirealism in cognitive science

Scientific realists argue that we are justified in committing towards our best scientific hypotheses in a rather strong sense, maintaining "a positive epistemic attitude towards the content of our best theories and models, recommending belief in both observable and unobservable aspects of the world described by the sciences" [9].

Such general attitude translates into more particular theses, depending on the specific variety of realism among the many currently discussed. Three of such theses, however, seem to be shared (in more or less qualified versions) by realists of all kinds [1, 29]. The first is the ontological thesis that the world that scientific theories aim to describe exists independently of our minds. The second is the epistemological thesis that knowledge of such mind-independent reality is indeed possible, being distilled in scientific theories and hypotheses that are at least approximately true representations of fragments of the world. The third is the methodological thesis that the methods adopted by scientists in their routine work do provide a means to

produce and accumulate such knowledge. While contemporary scientific antirealists tend to share the ontological thesis, they usually object both to the epistemological and the methodological one.

Given the very general nature of the above mentioned three theses, they apply in principle to any scientific discipline, including cognitive science and neuroscience in particular.[1] Still, as far as we know, a discussion of the realism/antirealism debate in this area is lacking at the moment.[2] One of the goals of the present paper is advancing such discussion. We don't aim at a comprehensive reconstruction of the possible realist and antirealist arguments that have been and could be advanced relative to the theories and methodologies underlying cognitive neuroscience; this would require a much longer and detailed work. Instead, we take a look at the history of cognitive science in general to briefly discuss the ontological and epistemological realist theses as applied to cognitive science (in this section) and then focus on the methodological thesis with reference to cognitive neuroscience in particular (in the next section).

As we suggest, one can read many episodes in the historical development of psychology and cognitive science as instances of the realism/antirealism debate. One has to do with the very birth of the cognitive science research program. During the first decades of the last century, psychology was largely influenced by behaviorism (for discussion see, e.g., [3]). According to that research program, any reference to internal (and hence unobservable) mental states should be removed from the vocabulary of psychology in order to study cognition in purely observable (behaviorial) terms. This radically empiricist view motivated a strong scepticism toward both the existence of mental entities and the possibility of scientifically investigate them. In this sense, behaviorism can be considered as an antirealist position, challenging both the ontological, the epistemological, and the methodological thesis of

[1] We follow here the traditional classification of neuroscience as one of the constitutive disciplines of cognitive science (along with philosophy, psychology, artificial intelligence, linguistics, and anthropology). For the purposes of this paper, we use "neuroscience" and "cognitive neuroscience" interchangeably, even if, strictly speaking, the latter is just a subarea of the former.

[2] As for neuroscience, this may well depend of the fact that this discipline is a young enterprise, dating back to the Nineties of the past century. Still, as we show in this section, the "pre-history" of cognitive neuroscience provides a number of interesting cues in the light of the realism/antirealism debate. For a quick presentation of the historical roots of contemporary neuroscience within traditional cognitive psychology and cognitive science, see [15]. For more general aspects see also [4, 23].

realism as applied to scientific psychology. As a consequence, one can interpret the "cognitive turn" that, between the Fifties and the Sixties, led to the birth of the classic cognitive science research program, as a realist response to the radically empiricist approach of behaviorism. Indeed, the crucial role played by mental representations and computations as the essential notions to investigate human and artificial minds apparently committed the newborn field of cognitive science to a more realist view of mental entities and cognition. Today, it seems fair to say that most cognitive scientists would subscribe to both the ontological and epistemological realist theses about mental entities and activities.

Of course, this is not to imply the the realism/antirealism debate in psychology and cognitive science has been settled in favor of the former position. Not only the controversy between cognitivists and behaviorists or neo-behaviorists continues today (see [28]), but also many other philosophical discussions over the past decades can be taken to represent a continuum of realist and antirealist positions about the nature of mental entities. For instance, the discussion about the existence of psychological states, especially those posited by folk psychology, provides a showcase of more or less realist positions, from dualism and epiphenomenalism (e.g., [26]), to functionalism and classical representationalism (e.g., [21]), reductionism, physicalism, and even more radical forms of eliminativism [11] (for discussion see also [42, 18]). A similar debate concerns the ontological status of phenomenal concepts or *qualia*: antirealists will maintain that *qualia* are mere illusions (e.g., [19]), or should be reduced to their representational and physical properties (e.g., [44]) or even eliminated (e.g., [12]), while realists tend to conceive *qualia* as irreducible and non-physical entities (see [10] for a general discussion).

Moreover, some developments within cognitive science have direct relevance to the realism/antirealism debate also in a different way. One example is the ongoing discussion about the "cognitive (im)penetrability" of perception [37]. A traditional view in both philosophy and psychology clearly distinguishes between perceptual experience, on the one hand, and cognition on the other hand. In this view, perception provides the real foundation for our knowledge of the external world: in particular, it is what we perceive that influences what we know about the world, and not the other way round. In this sense, human perceptual experience should be considered as independent from believes, desires, intentions, and cognition in general: in a slogan, perception is "cognitively impenetrable". This traditional view has been challenged by different authors (e.g., [13, 40]), who argue that our perceptual experience, such as, for instance, colour perception, is importantly affected both by cognitive states and by other theoretical constructs

(e.g., colour constancy) on which our cognitive system systematically relies. The debate about cognitive penetrability of perception has relevant consequences for different epistemological questions. First, these results put under scrutiny the role of observation as foundation for knowledge [43] and question the objectivity of empirical knowledge as evidence used to test scientific theories. Indeed, if perception is cognitively penetrated, the concepts we use to describe the external world might not represent it in a way that is objective and reliable, leading to a repertoire of conceptual tools that appears too epistemologically weak to inform our scientific theories. Second, as discussed by several scholars (e.g., [7, 5]), if observations are cognitively penetrated, that seems to point towards some strong claim about the theory-ladenness of empirical data and the incommensurability of scientific theories, in favor of an antirealist perspective.

Those mentioned are only few examples of the ontological and epistemological questions that may arise in cognitive science. In retrospective, it seems fair to say that, as far as the cognitivistic approach has become the dominant view in the field, a realist view of the ontological commitment of neuroscientific and psychological theories towards the unobservable entities posited by their hypotheses is the dominant one, while strongly antirealist positions seem less viable (for instance [18, 5]). In this connection, the debate seems to have moved from the ontological and epistemological level to the methodological one. In other words, assuming that unobservable cognitive entities and processes exist and can be accessed by scientific investigation, how reliable are the methods and techniques employed by cognitive (neuro)scientists in producing robust knowledge about them? In the next section, we address this question discussing some recent debates concerning experimental work based on fMRI in the field of cognitive neuroscience.

3 Reverse inference and empirical underdetermination

Since the last decade of the 20th century, neuroimaging techniques started having a greater and greater impact on the field of cognitive science. A remarkable role was played by fMRI, a non-invasive technique for studying the brain by analysing changes in the cerebral oxygenated blood flow as a proxy for neural activity. A crucial goal of fMRI-based experiments is to establish associations between the engagement of cognitive processes *Cog* (e.g., language, memory) and the activation of specific regions *Act* in the brain, as observed in fMRI data. Among the different inferential strategies em-

ployed by neuroscientists for reasoning about *Cog-Act* associations, reverse inference received increasingly critical attention, starting with the analysis advanced by neuroscientist Russell Poldrack in 2006 [33].

To illustrate, suppose we are interested in confirming the hypothesis advanced by some previous study that a certain brain region *Act*, namely, the *pars opercularis* within the inferior frontal gyrus of the left hemisphere of the human brain[3], has a pivotal role in language processing (*Cog*). Suppose further that, to test that hypothesis, we run an fMRI-based experiment involving a linguistic task on a group of subjects and the acquired brain data show a significant activation of *Act* (i.e., an increased average activation relative to a control group). Now, given that evidence and the hypothesis from previous literature about the *Cog-Act* association, the researchers might be confident in inferring the recruitment of *Cog* (viz. language processing) from the activity of *Act* (viz. the pars opercularis). In other words, they seem justified in performing the following reasoning (cf. [33]):

(P1) In the fMRI literature, when the cognitive process *Cog* was putatively engaged under the experimental task T, the brain area *Act* was active.

(P2) In the present fMRI study, under the experimental task T, the activation of brain area *Act* is observed.

(C) Therefore, in the present study, the cognitive process *Cog* under task T is engaged.

or, more formally:

(P1) $Cog \rightarrow Act$

(P2) Act

(C) Cog

Of course, this inference does not allow us to derive (C) with certainty, even when (P1) and (P2) are true. Indeed, from a logical viewpoint, reverse inference is a deductively invalid form of reasoning and, in particular, an instance of the "fallacy of affirming the consequent" [33]. This plain fact, by itself, is no decisive argument against the use of reverse inference (see [2] for critical discussion); still, it raises the question about the precise nature and

[3]Generally, when united to the region called *pars triangularis*, such area is known as Broca's area, from the pioneering studies of Paul Broca (1824-1880) investigating the role of this portion of the brain in elaborating language-related information.

role of such inferential pattern in neuroscientific studies. In this connection, the crucial issue is the lack of a unique, one-to-one association between *Cog* and *Act*, i.e., between the activity of a certain brain region *Act* and a unique cognitive function *Cog*. This well-known fact is known in the literature as the "problem of selectivity" of brain regions [33, 17]: in short, observing the activity of *Act* is not a sufficient and necessary condition for concluding that the function *Cog* was recruited. In fact, it is well-demonstrated from the fMRI literature that a variety of cognitive processes *Cog* may be responsible of the activity of the same area *Act*, leading to a many-to-one relationship between them.[4] The status of reverse inference has raised much discussion in recent years, with neuroscientists and philosophers both criticizing its uncontrolled use to derive bold conclusions from fMRI data and defending the underlying reasoning pattern as ultimately sound if used with caution [16]. Without entering the debate, here we would like to introduce a new way of looking at the problem of reverse inference, which is instrumental to our main goal, i.e., discussing the issue of realism within cognitive neuroscience. From this perspective, we think it is useful to interpret the problem, and the discussion it raised, through the lenses of what philosophers of science are used to conceptualize as a problem of empirical underdetermination (for a related discussion see also [23]).

Empirical underdetermination amounts to the fact that, at least in principle, it is always possible to find two or more different (and possibly incompatible) scientific theories or models which account for a given body of data. In this sense, empirical evidence cannot "determine" its own theoretical interpretation: in the presence of "empirically equivalent" theories or models, the data are just insufficient to establish which of them is preferable on purely empirical grounds. This may happen both because two (or more) theories are empirically equivalent in the strong sense of being in principle compatible with any possible body of data in the relevant field, and hence essentially indistinguishable [1]. Or, more modestly, because they are, at least for the time being, compatible with the same available evidence, so that a choice between them seems prima facie unwarranted.

[4]By saying that one or more cognitive processes may be responsible for, and hence explain, the activity of a certain brain region, we are not endorsing a specific view of explanation (such as a causal conception), nor entering in the debate about reductionism in neuroscience. The debate on reverse inference is essentially methodological, and has developed in the literature quite independently from other discussions about the nature of neuroscientific explanation, the issue of causation, and that of the relations between the mind and the brain. We thank an anonymous reviewer for prompting us to make this point clear.

As an example of the second kind drawn from the literature on cognitive science, let us consider the well-known "imagery debate" [45, 31]. The debate concerns the specific format of mental representations that allow us to imagine real-world objects when we do not directly perceive them. According to the "imaginistic" or "analogical" perspective (e.g., [25]) mental images are represented by our cognitive system in a depictive or iconic format, i.e., they basically operate as real pictures in our mind. On the contrary, the "symbolic" account [38] takes mental images to have the same representational format of sentences, that is, a symbolic or propositional one, with no role left for a properly imaginistic representation. In short, the core of the discussion focuses on whether "the concept of image can be used as a primitive explanatory construct" [36] or should be further analysed by means of lower-level, symbolic representations. Interestingly, the empirical evidence available when the debate arose (for instance [39, 25]), mainly based on behavioral experiments, was compatible with both perspectives. In other words, it was hard to establish which of the two rival accounts, i.e., the analogical or the symbolic account, was preferable as a theory of the representation format of mental images.[5]In this sense, one can see the early imagery debate as a case of empirical underdetermination, where two competing theoretical accounts are empirically equivalent with respect to the available evidence, so that a rational choice between them is hard to justify on empirical grounds.

Going back to reverse inference, it seems to us that this kind of inference raises a clear case of underdetermination for hypotheses concerning the engagement of different cognitive processes. Of course, we are talking here of a quite modest form of underdetermination relative to low-level hypotheses of the form "in the present case, cognitive process Cog is engaged". As said, the typical issue of reverse inference arises when two distinct cognitive processes Cog_1 and Cog_2 are both associated to a certain area Act which is activated in a fMRI experiment. If our goal is explaining such activation by pointing to the engagement of one of the two processes, we face a problem of underdetermination. In other words, the schema introduced above can be revised as follows:

(P1) In the fMRI literature, when the cognitive process Cog_1 was putatively

[5]Nowadays, the debate appears resolved in favor of those advocating the depictive format of mental images. Indeed, as noticed by [31], there is strong evidence (based on neuroimaging works such as [32]) in favor of the thesis that humans represent information in multiple ways, with depictive images playing a key role in many areas of cognition.

engaged under the experimental task T, the brain area Act was active.

(P2) In the fMRI literature, when the cognitive process Cog_2 was putatively engaged under the experimental task T, the brain area Act was active.

(P3) In the present fMRI study, under the experimental task T, the activation of brain area Act is observed.

(C) Therefore, in the present study, the cognitive process Cog_1 (instead of Cog_2) under task T is engaged.

Since both Cog_1 and Cog_2 may be responsible of the activity of Act, Act may equally support or confirm the recruitment of both Cog_1 and Cog_2, respectively, so making the conclusion in favor of (say) Cog_1 apparently untenable. Thus, as far as Cog_1 and Cog_2 are considered as competing neuroscientific hypotheses about the possible causes of Act, they are compatible with the same data and hence underdetermined by the available experimental evidence. In this sense, the problem of underdetermination – here depending on the problem of selectivity – seems intrinsic to the methodology of cognitive neuroscience as far as reverse inference is concerned. Whether such problem justifies a skeptical view of the neuroscientific quest for explanatory cognitive hypotheses of neural activity is discussed in the next section.

4 A realist look at the problem of reverse inference

In the last section, we argued that a problem of underdetermination lies at the methodological core of cognitive neuroscience, arising, in principle, each time reverse inference is employed to derive hypotheses about the engagement of cognitive processes from fMRI data. Since underdetermination figures prominently in many arguments against scientific realism in general, this motivates a reflection on whether the problem of reverse inference justifies an antirealist attitude about neuroscientific hypotheses and explanations. In this section, we argue that this is not the case.

Let us first recall how underdetermination seems to defy a realist view of scientific methodology in general. The intuition is that, if two competing theories or hypotheses are actually equivalent as far as the available evidence is concerned, a rational, evidenced-based choice between them must be unwarranted. In fact, given that this choice is underdetermined by empirical data, it is unclear what should justify a preference of one hypothesis over the other. In particular, if the two hypotheses differ in the unobservable

constructs they posit (for instance, because they assume the engagement of different cognitive processes), one has no reason to think, on the basis of the available evidence, that one may be a better approximation to the truth than the other. In turn, this undermines the realist's claim that we should trust our better supported hypotheses as approximate representations of the target domain [9].

The general argumentative strategy just presented has long been used in the philosophical discussion to advance more specific arguments against different realist theses, often based on historical case studies. Of course, realists have reacted in a number of ways to each such critique, and the literature is now plenty of arguments and counterarguments surrounding the notion of underdetermination. Without trying to assess the debate, let us mention two of the main strategies employed by realists to resist antirealist conclusions from empirical underdetermination (for a more extended discussion, see [1]). First, one may doubt whether genuine cases of strong empirical equivalence of two actually different theories may exist, both from a conceptual and a historical point of view. If not, one may argue that theories that seem to be genuinely equivalent are actually two different versions of the same theory, so that "underdetermination does not look like a concrete problem in scientific practice, but a merely in-principle risk not to be taken too seriously" [1]. In that case, the antirealist argument based on underdetermination would be a non-starter. Second, even acknowledging that two genuinely different theories are compatible with the same evidence, this doesn't mean, the realist counter-objection goes, that they are empirically equivalent in any strong sense. In fact, the evidence in favor of a theory doesn't depend only on its empirical performance (so to speak), but also on the "theoretical virtues" enjoyed by theories, like for instance simplicity, high prior probability, explanatory power, and the like. If this is true, theoretical choice remains possible also in the face of empirical underdetermination. In sum, there is not shortage of realist answers to the challenge based on underdetermination, even if, as it is to be expected, antirealists are no less prolific in devising new objections.

Interestingly, the recent debate on the issue of reverse inference in cognitive neuroscience has apparently followed in the steps of the philosophical debate on underdetermination and its consequences for realism. This confirms our suggestions that looking at reverse inference as a problem of empirical underdetermination provides a useful entry point to the realism/antirealism debate within cognitive (neuro)science. To support this claim, we briefly discuss two cases in which the way reverse inference has been discussed is very close to how a defence of a realist stance in the face of underdetermination

occurs in the philosophical literature.

Reverse inference and Bayesian confirmation. Starting with Poldrack [33], neuroscientists have resorted to a Bayesian analysis to address the problem of reverse inference and provide a conceptual analysis of such pattern of reasoning. Let consider, for instance, two competing cognitive hypotheses Cog_1 and Cog_2 and evidence Act provided by the activation of some brain area as observed with fMRI. Proponents of the Bayesian analysis of reverse inference claim that the rational evaluation of Cog_1 and Cog_2 should be guided by their respective probability, as estimated via Bayes' rule. Thus, if the posterior probability of Cog_1 given Act, calculated as

$$p(Cog_1|Act) = \frac{p(Act|Cog_1)p(Cog_1)}{p(Act)} \tag{1}$$

is greater than the posterior probability of Cog_2 (calculated in the same way as above), there is reason to prefer Cog_1 over Cog_2 as the most credible hypothesis accounting for the observed evidence Act. In this way, reverse inference is formalized as Bayesian reasoning, thus providing a possible way of choosing between Cog_1 and Cog_2 the best hypothesis given the available evidence Act.

Interestingly, the above proposal is in line with a classical argument in philosophy of science advanced by realists proposing a Bayesian response to the issue of empirical underdetermination. According to them, even if two competing hypotheses are empirically equivalent, it is still possible in general to differently assess them in Bayesian terms [22, 35, 1]. Indeed, even if the likelihoods of the two hypotheses — in our case, $p(Act|Cog_1)$ and $p(Act|Cog_2)$ — may be the same, expressing the fact that they are empirically equivalent, i.e., explain Act equally well, their posteriors could well be different, if Cog_1 and Cog_2 differ in their prior probabilities $p(Cog_1)$ and $p(Cog_2)$. In general, such prior probabilities depend on how likely scientists deem the relevant hypotheses before any particular evidence is considered; in turn, this will depend on relevant background knowledge about the two hypotheses, including an assessment of their theoretical virtues as discussed in Section 4. In this way, a Bayesian analysis allows the realist to discriminate between empirically equivalent hypotheses, at least as far as their priors differ, thus overcoming the issue of empirical underdetermination.

In this connection, however, one should note that actual implementations of such Bayesian analysis of reverse inference in current neuroscientific practice cannot sustain such realist reading of reverse inference. The plain

reason is that all proposals along these lines assume that the prior probabilities of the relevant hypotheses, for instance $p(Cog_1)$ and $p(Cog_2)$ are equal [33, 24].[6] This has motivated non-Bayesian proposals, like the one advanced by Edouard Machery [27], to treat reverse inference in purely "likelihoodist" terms, i.e., assessing hypotheses like Cog_1 and Cog_2 only in terms of their likelihoods $p(Act|Cog_1)$ and $p(Act|Cog_2)$. While defensible on their own, such proposals cannot of course be employed by realists to defend neuroscientific methodology in the face of underdetermination (nor this is their purpose); in fact, for empirically equivalent hypotheses it is sensible to assume that their likelihoods are the same, and hence cannot guide theoretical choice.

This, however, is not the end of the story. As we suggested elsewhere [14], one can defend a more sophisticated analysis of reverse inference, based on the notion of Bayesian confirmation, according to which the choice between Cog_1 and Cog_2 is not guided by a plain assessment of their posterior probabilities in the light of Act, but by the degree of empirical support or confirmation assigned to each hypothesis on the basis of Act. Such move allows one to differently evaluate empirically equivalent cognitive hypotheses by choosing the one with the highest degree of empirical support, even when one assumes that their prior probabilities are equal. This is especially true for those measures of Bayesian confirmation (like the so-called likelihood ratio) that only depend on the likelihoods $p(Act|Cog_1)$ and $p(Act|Cog_2)$ of the competing hypotheses, that can be empirically evaluated on a case-by-case basis. Interestingly, Poldrack himself proposes to use the so-called Bayes factors of competing hypotheses Cog_1 and Cog_2 in order to empirically compare them in the context of reverse inference. Since the Bayes factor, widely employed in a number of scientific disciplines, is a confirmation measure in the philosopher's sense, and in particular it is essentially equivalent to the likelihood ratio measure [41], this further illustrates how the current debate on reverse inference seems to move along the lines already traced by the realims/antirealism debate in the philosophy of science.

[6] Flat priors are assumed for different reasons. For instance, neuroscientists might have no specific expectation about the recruitment of a certain process during a task. Moreover, flat priors are generally assumed to avoid selection biases affecting the published literature. This assumption is especially true for the most widely employed tool for performing automated reverse inference based on large fMRI databases, i.e., NeuroSynth [46]. For a discussion of NeuroSynth in the light of such methodological issue see [15].

Reverse inference as inference to the best explanation. The Bayesian analysis outlined above is not the only way that philosophers and neuroscientists have tried to make sense of reverse inference and its use in experimental practice. Other scholars have proposed to interpret reverse inference as a form of abductive reasoning in the sense of Peirce (again, Poldrack himself mentions in passing such interpretation in his seminal paper [33]). According to such proposals [6, 34, 8], reverse inference can be viewed as an inference from the observed effect *Act* to the engagement of some cognitive process *Cog* as its putative cause. Following the terminology of [8], such pattern of reasoning can be employed both in a "weak" and in a "strong" form. In weak reverse inference, the conclusion about the engagement of *Cog* is interpreted, in a more heuristic way, as a suggestion of a possible cognitive explanation of the observed activation, to be subjected to further exploration and testing. In its strong reading, instead, one views reverse inference as an "inference to the best explanation" (IBE) of the available evidence *Act*, which then confirms its conclusion *Cog* as the most plausible candidate hypothesis accounting for the experimental results.

The discussion on the prospects and relative merits of weak and strong reverse inference is still open [8, 16]. To our purposes, it will be sufficient to note that abductive reasoning, and especially IBE, play a prominent role in the realism/antirealism debate, on at least two levels. First, the "no miracle argument", the central and perhaps more powerful argument in favor of scientific realism, can be construed as an abductive argument inferring, from the observed success of a scientific theory, to its approximate truth as the best explanation of such success [1, 9, 30]. Second, discussions of IBE crucially rely on the analysis of different theoretical virtues enjoyed by different hypotheses [20] and, as recalled in Section 4, such virtues play an important role also in realist arguments concerning underdetermination. Interestingly, IBE has recently entered the debate on reverse inference exactly along these latter lines.

As recent work has highlighted [8, 16], the evaluation of competing cognitive hypotheses in the light of available evidence from fMRI studies is not limited to the fact that the relevant hypotheses "fit the data", meaning they can account for the observed neural activations. Instead, other "virtues" of such hypotheses are assessed, like how well they cohere with currently accepted neuroscientific knowledge; how "good" they are as potential explanation of the evidence and how better they perform, under this respect, relative to their competitors; how able they are to unify different kinds of evidence (e.g., neural, behavioral, derived from animal models, etc.); and so on. Again, all such virtues have been carefully analyzed in the philosophi-

cal discussion of IBE, highlighting how simplicity, non-*adhocness*, coherence with background knowledge, greater explanatory power over closer competitors, unifying power, and the like all contribute in determining, first, what a "best explanation" is and, second, how reliable an IBE is in favor of the hypothesis best performing on such criteria [20]. In short, to the extent to which reverse inference can be construed as an instance of IBE and such kind of inference allows to overcome some problems connected with empirical underdetermination, the analysis just outlined suggests how a realist view of current scientific practice is tentatively warranted in the field of cognitive neuroscience.

5 Concluding remarks

In this paper, we looked at the issue of reverse inference in cognitive neuroscience through the lenses of the philosophical discussion of empirical underdetermination as a problem for a realist view of scientific inference. We first argued that the problem of reverse inference is essentially one of empirical underdetermination, thus challenging the methodological thesis of scientific realism, according to which the methods routinely employed by working scientists provide effective means to accumulate genuine knowledge about the relevant domain. Against this background, the critiques of reverse inference raised in the literature may apparently favor an antirealist view of neuroscientific methodology, even if leaving perhaps untouched the ontological and epistemological status of cognitive entities in general.

We then discussed two trends in the recent debate on reverse inference: one favoring a Bayesian analysis and the other considering reverse inference as IBE. In both cases, neuroscientists and philosophers have put forth arguments for favoring one cognitive hypothesis over others despite the problem of selectivity. As we suggested, these arguments amount to providing strategies for dealing with the issue of cognitive hypotheses which are empirically equivalent relative to the available neural evidence. Interestingly, the debate on reverse inference appears to closely track that on realism/anti-realism in general: the principal arguments in favor of reverse inference largely rely on theoretical virtues that competing but empirically equivalent hypotheses may possess differently.

In our view, this provides an intriguing instance of current scientific practice illustrating some more general philosophical principles at work. As we argued, the ongoing discussion on reverse inference, first, casts doubt on the notion that the problem of empirical underdetermination necessarily

licenses an anti-realist attitude to (neuro)scientific methodology; second, it suggests that a realist view of reverse inference and related methodological problems is at least tenable, and in fact embedded in two central arguments favoring the employment of such inference.

References

[1] Alai, M. (2017). The debates on scientific realism today: knowledge and objectivity in science. In: E. Agazzi (ed), *Varieties of Scientific Realism*, Springer, 19–47.

[2] Anderson, M. L. (2010). Review of *Neuroeconomics: Decision making and the brain*, *Journal of Economic Psychology* 31: 151-–154.

[3] Bechtel, W., Abrahamsen, A., Graham, G. (2017). The life of cognitive science. In: G.Graham, W. Bechtel (eds),*A companion to cognitive science*, Blackwell Publishing Ltd., 1–104.

[4] Bechtel, W., Huang, L. T. (2022). *Philosophy of neuroscience*, Cambridge, Cambridge University Press.

[5] Beni, M. D. (2021). Cognitive Penetration and Cognitive Realism, *Episteme*: 1–16.

[6] Bourgeois-Gironde, S. (2010). Is neuroeconomics doomed by the reverse inference fallacy?,*Mind & Society* 9 (2): 229–249.

[7] Brewer, W. F., Lambert, B. L. (2001). The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process, *Philosophy of Science* 68 (S3): S176–S186.

[8] Calzavarini, F., Cevolani, G. (2022). Abductive reasoning in cognitive neuroscience: weak and strong reverse inference, *Synthese* 200 (2): 1–26.

[9] Chakravartty, A. (2017). Scientific Realism. In: E. N. Zalta (ed), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

[10] Chalmers, D. J. (2006). Phenomenal concepts and the explanatory gap. In: T. Alter, S. Walter (eds), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.*

[11] Churchland, P. M. (1981). Eliminative materialism and propositional attitudes, *The Journal of Philosophy* 78 (2): 67–90.

[12] Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states, *The Journal of Philosophy* 82 (1): 8–28.

[13] Churchland, P. M. (1988). Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor, *Philosophy of science* 55 (2): 167–187.

[14] Coraci, D., Cevolani, G. (2022). L'analisi bayesiana dell'inferenza inversa in neuroscienza: una critica, *Sistemi intelligenti* 34 (2):209–234.

[15] Coraci, D., Calzavarini, F., Cevolani, G. (2023). Reverse Inference, Abduction, and Probability in Cognitive Neuroscience. In: L. Magnani (ed), *Handbook of Abductive Cognition*, Springer.

[16] Coraci, D., Cevolani, G., Douven, I. *Inference to the Best Neuroscientific Explanation*, manuscript.

[17] Del Pinal, G., Nathan, M. J. (2017). Two kinds of reverse inference in cognitive neuroscience, *The human sciences after the decade of the brain*, Elsevier: 121–139.

[18] Demeter, T. (2009). Two kinds of mental realism, *Journal for general philosophy of science* 40 (1): 59–71.

[19] Dennett, D. C. (1988). Quining qualia, *Consciousness in contemporary science*: 42–77.

[20] Douven, I. (2022). *The art of abduction*, Cambridge (MA), MIT Press.

[21] Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*, MIT press.

[22] Glymour, C. (1980). Theory and evidence. In: Hanson, N. R., C W. Humphreys (eds), *Perception and discovery*, Springer.

[23] Hanson, S. J. E., Bunzl, M. E. (2010). *Foundational issues in human brain mapping*, Cambridge (MA), MIT Press.

[24] Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data, *NeuroImage* 84, 1061–1069.

[25] Kosslyn, S. M. (1980). *Image and mind*, Harvard: Harvard University Press.

[26] Lyons, J. C. (2006). In defense of epiphenomenalism, *Philosophical Psychology* 19 (6): 767–794.

[27] Machery, E. (2014). In Defense of Reverse Inference, *The British Journal for the Philosophy of Science* 65 (2): 251–267

[28] Nanay, B. (2019). Entity realism about mental representations, *Erkenntnis*: 1–17.

[29] Niiniluoto, I. (1999). *Critical scientific realism*, Oxford, Oxford University Press.

[30] Niiniluoto, I. (2018), *Truth-Seeking by Abduction*, Springer.

[31] Pearson, J., Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate, *Proceedings of the national academy of sciences* 112 (33): 10089–10092.

[32] Pearson, J., Naselaris, T., Holmes, E. A., Kosslyn, S. M. (2015). Mental imagery: functional mechanisms and clinical applications, *Trends in cognitive sciences* 19 (10): 590–602.

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

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

[35] Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*, London-New York, Routledge.

[36] Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery, *Psychological bulletin* 80 (1): 1–24.

[37] Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science, *Behavioral and Brain sciences* 3 (1): 111–132.

[38] Pylyshyn, Z. W. (1981). The imagery debate: Analogue media versus tacit knowledge, *Psychological review* 88 (1): 16–45.

[39] Shepard, R. N., Metzler J. (1971). Mental rotation of three-dimensional objects, *Science* 171 (3972): 701–703.

[40] Siegel, S. (2019). Cognitive penetrability and perceptual justification, *Contemporary Epistemology: An Anthology*: 164–178.

[41] Sprenger, J., Hartmann, S. (2019). *Bayesian philosophy of science*, Oxford University Press.

[42] Sprevak, M. (2017). Realism about cognitive science. In: J. Saatsi (ed.) *Routledge Handbook of Scientific Realism*, London, Routledge: 357–368.

[43] Stokes, D. (2013). Cognitive penetrability of perception, *Philosophy Compass* 8 (7): 646–663.

[44] Tye, M. (1997). *Ten problems of consciousness: A representational theory of the phenomenal mind*, MIT press.

[45] Tye, M. (2000). *The imagery debate*, Cambridge (MA), MIT Press.

[46] Yarkoni, T., et al. (2011). Large-scale automated synthesis of human functional neuroimaging data, *Nature methods* 8 (8): 665–670.