Synthese (2022) 200:70
https://doi.org/10.1007/s11229-022-03585-2

ORIGINAL RESEARCH

Abductive reasoning in cognitive neuroscience: weak and strong reverse inference

Fabrizio Calzavarini[1,2] · **Gustavo Cevolani**[2,3]

Received: 30 April 2021 / Accepted: 20 January 2022 / Published online: 5 March 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract
Reverse inference is a crucial inferential strategy used in cognitive neuroscience to derive conclusions about the engagement of cognitive processes from patterns of brain activation. While widely employed in experimental studies, it is now viewed with increasing scepticism within the neuroscience community. One problem with reverse inference is that it is logically invalid, being an instance of abduction in Peirce's sense. In this paper, we offer the first systematic analysis of reverse inference as a form of abductive reasoning and highlight some relevant implications for the current debate. We start by formalising an important distinction that has been entirely neglected in the literature, namely the distinction between weak (strategic) and strong (justificatory) reverse inference. Then, we rely on case studies from recent neuroscientific research to systematically discuss the role and limits of both strong and weak reverse inference; in particular, we offer the first exploration of weak reverse inference as a discovery strategy within cognitive neuroscience.

Keywords Reverse inference · Abduction · Cognitive neuroscience · Justification · Discovery

1 Introduction

Abductive inference is reasoning backwards from facts to their possible explanations, or from effects to their possible causes. It is at play in a wide array of contexts, from science to everyday life; for instance, when we infer that it rained since the grass is wet, or when a doctor diagnoses a strep throat from fever and white spots on the

✉ Fabrizio Calzavarini
fabrizio.calzavarini@unibg.it

1 Department of Letter, Philosophy, Communication, University of Bergamo, Via Pignolo 123, Bergamo, Italy

2 Center for Logic, Language, and Cognition, Turin, Italy

3 IMT School for Advanced Studies Lucca, Piazza S. Francesco, 19, 55100 Lucca, Italy

patient's tonsils. Starting at least with Peirce, philosophers studied abduction—often under the label "inference to the best explanation"—both from a logical point of view and in connection with the methodology of science and of different expert practices. Despite a decade-long discussion, no consensual view of the nature, role, and significance of abductive reasoning has emerged so far. Disagreement has to be registered at various levels. One central debate concerns whether abduction has a mainly heuristic function—that of generating new explanatory hypotheses and assisting discovery—or also a justificatory role, one of evaluating and possibly accepting selected hypotheses. Moreover, some authors even doubt that abduction is really important or needed in ordinary and scientific inference (e.g., Norton, 2016), and question the idea that explanatory considerations can have a place within ordinary Bayesian confirmation theory (for relevant discussion, see, e.g., Niiniluoto, 1999, 2018; Lipton, 2004; Douven & Schupbach, 2015; Schurz, 2017).

Quite unrelatedly to this philosophical discussion, in the last years working scientists have developed various methods to successfully deal with abductive inference in a variety of fields, from medical diagnostics to evolutionary theory, to Artificial Intelligence (e.g., Niiniluoto, 2011, 2018; Schurz, 2017). In this paper, we focus on abductive reasoning as performed in cognitive neuroscience, i.e., the study of the biological especially neural processes that underlie cognition and mental activities. In this area, functional resonance magnetic imaging (fMRI) plays a crucial role in the exploration of brain activity. This technique is being used in two different ways. First, neuroscientists build brain maps by studying which regions are activated by different mental processes (as elicited by different tasks, e.g., face recognition or language processing). This is so-called *forward inference*, from mental processes to their putative neural correlates. Second, researchers routinely employ the inverse reasoning strategy, inferring from specific activation patterns to the engagement of particular mental processes. This so-called *reverse inference* plays a crucial role in many applications of fMRI, both inside and outside cognitive neuroscience. These include the diagnosis of disorders in patients with acquired brain pathologies such as schizophrenia and Alzheimer's disease (Costa et al., 2021), the well-known experimental studies of moral reasoning as pioneered by Greene et al. (2001), and most studies in so called neuroeconomics (Bourgeois-Gironde, 2010).

In recent years, reverse inference has attracted a great deal of attention, especially after neuroscientist Russell Poldrack (2006) denounced an uncontrolled "epidemic" of this reasoning pattern, cautioned against its (improper) use and pointed to its crucial weakness. In further work, Poldrack and collaborators applied machine learning and data mining techniques to automatically explore big fMRI data sets to extract relevant correlations between mental processes and activation patterns to be used in making reverse inference more robust and reliable (the *NeuroSynth* project, see Yarconi et al., 2011). The debate is still open, and the present methodological status of reverse inference is highly controversial (Glymour & Hanson, 2016; Hutzler, 2014; Machery, 2014; Nathan & Del Pinal, 2017; Poldrack, 2008, 2011; Weiskopf, 2020). Interestingly, Poldrack himself noted in passing that reverse inference could be analysed as an instance of abductive reasoning, but neither he nor others developed further this suggestion.

In this paper, we offer the first systematic analysis of reverse inference as a pattern of abductive reasoning. Our central claim is that the first step towards bringing some

order to the discussion is to formalise an important distinction that has been entirely neglected in the literature on the methodology of cognitive neuroscience, namely the distinction between weak (strategic) and strong (justificatory) abduction. Emphasising this distinction allows us to clearly separate different forms of reverse inference—to which we shall refer to as "weak reverse inference" and "strong reverse inference", respectively—that are usually conflated in that literature. As we argue, this theoretical move has two main benefits. First, it allows for a better understanding of how neuroscientists use fMRI data, as well as a better reconstruction of their reasoning strategies and methods. Second, it helps in clarifying current debates among cognitive neuroscientists, who may gain a better appreciation of the potentialities and limits of reverse inference, and hence improve their theoretical and experimental practices.

Overall, our discussion provides the first attempt to systematically apply the theoretical and conceptual tools developed in the philosophical study of abduction to the analysis of reverse inference in cognitive neuroscience, thus bridging two different kinds of literature which have been so far largely independent. We proceed as follows. In the first section, we explore the distinction between weak and strong abduction as it has been discussed in philosophy. In the second section, we discuss the attempts made so far to analyse reverse inference in cognitive neuroscience and we introduce the distinction between strong and weak reverse inference. In the third and the fourth sections, we rely on case-studies from recent neuroscientific research to systematically explore the role of both strong and weak reverse inference, and we offer the first exploration of weak reverse inference as a discovery strategy.

2 Weak and strong abductive inference

Peirce called "abduction" the pattern of reasoning—for which he also used the terms "retroduction" (CP 1.68) or "hypothesis" (CP 2.623)[1]—involved in «the operation of adopting an explanatory hypothesis» for a given piece of evidence (CP 5.189). For instance: «Fossils are found; say, remains like those of fishes, but far in the interior of the country. To explain the phenomenon, we suppose the sea once washed over this land» (CP 2.625). Peirce clearly saw that, even if the truth of the premises is taken for granted, the conclusion of an abductive argument may be false: in other words, like induction, and contrary to deduction, abduction is a form of ampliative and uncertain reasoning. The logical form of an abductive inference, according to Peirce (CP 5.189), is the following:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.

Clearly, the inference "if A then C; but C; therefore, A" is deductively invalid, being an instance of the fallacy of "affirming the consequent". However, Peirce noted that a scientific argument can be logically invalid but still effective in making its conclusion worth of further consideration (CP 5.192). Accordingly, although their conclusions

[1] When quoting from Peirce's Collected Papers (CP, Hartshorne et al. 1931–1958), we follow the convention of citing the number of the volume followed by the number of the relevant paragraph.

are always tentative and conjectural, Peirce argued that abductive arguments provide a fundamental form of inference both in scientific and everyday reasoning.

During the 1960s, abduction started attracting systematic attention from philosophers of science. Hanson (1958) suggested that Peirce's schema provides "a logic of scientific discovery" and Harman (1965) argued that "inference to the best explanation" (IBE, for short), as he called abduction, is the core of any ampliative or non-deductive inference. These pioneering contributions made clear that there are at least two different ways –respectively, a "weak" and a "strong" one—of assessing the proper role of abductive inference. According to the first, weak interpretation, abduction has a primary discovery (or "strategic" or "heuristic", see Schurz, 2008, p. 203) function, that of suggesting or finding promising or "test-worthy" hypotheses which are then set out to further inquiry or empirical testing. According to the second, strong (or justification) reading, abduction can be formulated as a rule of acceptance, since it gives reasons to tentatively accept its conclusion as the "best" explanatory hypothesis among the available ones.

In different writings, Peirce seems to endorse the weak or the strong view of abductive inference, or both. In the last decades, there has been a lot of discussion about the proper interpretation of abduction, both within Peircean scholarship and within (formal) philosophy of science. On the one hand, contemporary defenders of IBE have tended to see Peircean abduction as a way of justifying an explanatory hypothesis (see Mcauliffe, 2015). In his influential book *Inference to the Best Explanation* (2004, pp. 56–57), for example, Peter Lipton bluntly claims that Peircean abduction is a conceptual precursor of IBE. On the other hand, it is clear that Peirce also considered weaker forms of abduction, which will be relevant for our discussion in the following. According to one of these weaker readings, abduction should be construed as a discovery procedure whose main function is to generate, but not justify, novel hypotheses (see Minnameier, 2004; Campos, 2011). Abduction, according to Peirce, «strikes out a new suggestion», and is «the only logical mechanism which introduces a new idea» (CP 5.590). He also emphasises that abductive conclusions are not always invented *ex novo* but can have various degree of creativity. As noted by Anderson (1987), Peirce actually distinguishes two kinds of abductive novelty, *rearrangement* and *concept creation*: «[t]he first kind of novelty […] is a combination which is different from past views, but which is grounded in ideas or perceptions we have already. […] The second grade of novelty, which is not always easy to distinguish from the first, is the creation of a new concept—that is, of an idea which we have not previously had» (pp. 47–48).

A second function of weak abduction defended by Peirce is not *generating* hypotheses but rather *selecting* which of the potential explanations of a given phenomenon are worthy candidates for further investigation (see Frankfurt, 1958; McKaughan, 2008). According to this conception, abduction again does not lend any support or justification to a hypothesis. As Douven (2017) observes, «[o]n this understanding, abduction could still be thought of as being part of the context of discovery. It would work as a kind of selection function, or filter, determining which of the hypotheses that have been conceived in the stage of discovery are to pass to the next stage and be subjected to empirical testing». Interestingly, Peirce also notes that sometimes we should prefer *uberty* over *security*, selecting those hypotheses that are risky and a priori unlikely

but have the potential to open new paths of research that might be fertile (CP 8.384, 1913).

Apart from discussion within Peircean scholarship, and quite independently from it, abduction has been the focus of an intense philosophical debate over the years—so intense that Hintikka (1998) famously claimed that this is «the fundamental problem in contemporary epistemology». Most philosophers have followed Harman in equating abduction with IBE, discussing it mainly in the context of justification. Is abduction a reliable method of confirmation? How can abductive arguments be improved? The "textbook version of abduction" (Douven, 2017) is something along the following lines: «[g]iven evidence E and candidate explanations $H_1,\ldots H_n$ of E, infer the truth of *that* H_i which best explains H». This basic formulation raises a number of critical issues, having to do with the correct explications of the notions of candidate explanations and best explanation. Just to mention one, in the above formulation the notion of "best explanation" is always relative to a set of *available* explanations, which is inevitably restricted at least by the scientists' imagination and other contingencies. It is thus possible that the explanation that is actually the *best* is included within the set of those explanations that scientists have not considered due to lack of imagination, time, or other reasons. In short, it is possible abductive reasoning leads us to believe the "best of a bad lot" (van Fraassen, 1989; see Douven, 2017 for discussion and Schupbach, 2014 for a rebuttal of van Fraassen's argument).

Another critical issue about the justificatory status of abduction has to do with the criteria for deciding which is the *best* among the alternative candidate explanations (cf. Lipton, 2004). One immediate suggestion is to identify the best explanation with that hypothesis H_i that is most probable given the evidence E (i.e., it has the highest degree of *posterior probability*), or that is most strongly supported by E (i.e., it has the highest degree of *confirmation*). In this sense, as some scholars have suggested (Salmon, 2001), abduction can be formalised using Bayes theorem or one of the measures of probabilistic support studied within Bayesian confirmation theory (see Crupi, 2020; Niiniluoto, 2018). However, the relation, and even the compatibility, between abduction and Bayesian reasoning is quite controversial, for there is no direct and clear connection between probabilistic and explanatory considerations in assessing hypotheses (see Douven, 2017 and Sprenger & Hartman, 2019 for discussion). Thus, some scholars have even argued that either abduction is reducible to Bayesian theory, or it is epistemically irrelevant (Roche & Sober, 2013). Other scholars reached similar pessimistic conclusions about abduction based on different considerations, such as its weak evidential role (van Fraassen, 1989) or its lack of unity (Norton, 2016). According to Norton (2016), for example, abduction as IBE is an «overrated argument form» (p. 200).

At the same time, other philosophers have followed Hanson in exploring the weak functions of abduction in the context of discovery rather than justification. As observed by Paavola (2006), «[i]n Hanson's view, abduction is a weak form of inference that relates to the first phase of inquiry. This "weakness" means that abduction is supposed to give plausible candidate hypotheses, not necessarily true explanations, which then have to be verified and tested by other means» (p. 97). Magnani (2001, 2009) discusses a similarly weak conception of abduction, distinguishing *creative* abduction (abduction that generates new hypotheses) and *selective* abduction (abduction that merely selects

from a gamut of pre-stored hypotheses). Selective abduction in Magnani's terms should not be confused with IBE, for «all we can expect of our "selective" abduction, is that it tends to produce hypotheses for further examination that have some chance of turning out to be the best explanation» (2009, p. 97–98). Similarly, in discussing his fine-grained taxonomy of patterns of abduction (2008; see also 2017), Schurz advocated weak abduction by arguing that the «crucial function of a pattern of abduction […] consists in its function as a *search* strategy which leads us, for a given kind of scenario, in a reasonable time to a most promising explanatory conjecture which is then subject to further test» (2008, p. 205).

In sum, although the distinction between weak and strong abduction is usually acknowledged in the literature, it only plays a marginal role in most discussions. The reason is that many contributors to the debate tend either to acknowledge only one of the two concepts as legitimate, thus discarding, more or less explicitly, the other one; or to propose reconstructions of abduction that attempt to incorporate both its weak and strong form within one single model, with the risk of conflating them.[2]

Admittedly, a categorical differentiation between weak and strong abduction is probably not possible, nor useful, since, as for instance Niiniluoto (1999, p. 442) points out, «abduction as a motive for pursuit cannot always be sharply distinguished from consideration of justification». Indeed, several scholars doubt that questions of justification can be neatly separated from questions of discovery in general (see, e.g., Schickore, 2018): as Schurz (2008) puts it, «[a]ll inferences have a justificational (or 'inferential') and a strategical (or 'discovery') function, but to a different degree» (p. 203). For instance, selective abduction is defended by Magnani (2001) as a procedure for discovery; however, justificatory considerations clearly intervene when competing hypotheses are assessed as more or less plausible given the available evidence. While we concur that neatly separating the strategic and the justificatory function of concrete cases of abductive inferences in science is possible, we also believe that the two moves mentioned above (overlooking one of the two functions and trying to unify them) are unwarranted. In particular, we believe that the distinction between weak and strong abduction highlights a clear, underlying contrast between two fundamental functions of abductive reasoning, which are both legitimate objects of study and should be carefully distinguished in the philosophical analysis. Thus, we depart here both from fully "compatibilist" accounts of abduction—which tend to conflate the two functions of abduction into a single concept—and by "unilateral"

[2] To mention but a few examples of the first tendency, the entry on "Abduction" in the *Stanford Encyclopaedia of Philosophy* by Douven (2017) only focuses on the "modern" (i.e., strong) sense of abduction, IBE, confining the discussion of the "historical" (i.e., weak) sense to a short supplement. On the opposite side, scholars such as Minnemaier (2004), Campos (2011), and Mcauliffe (2015) have argued against the tendency to equate Peircean (weak) abduction and IBE, claiming that only the first concept can be legitimately called "abduction". As for the second path mentioned above, for instance Lipton argues for a «version of IBE thus includes two filters, one that selects plausible candidates, and a second that selects from among them» (2004, p. 64) as a unified model of weak ad strong abduction. Similarly, Schurz (2017) explicitly equates abduction with IBE (p. 152) but, at the same time, he carefully analyses the strategic function of abduction, concluding that «the justificatory function of abduction is minor» (p. 153). A minority of scholars avoids overlooking the distinction between weak and strong abduction. For instance, Paavola (2004) explicitly distinguishes between what he calls "Hartmanian abduction" (IBE) and "Hansonian abduction" (weak abduction) even if, following Lipton (2004), he then discusses Hartmanian abduction more as a "method of discovery" than as an instrument for justification.

accounts—which try to discount one of the two functions as immaterial or irrelevant. To be sure, it is not the primary aim of this paper to defend the distinction between strong and weak abduction on theoretical grounds. Instead, in the following, we shall attempt to show the usefulness of such a distinction within a specific field of scientific research, namely cognitive neuroscience.

3 Reverse inference as abductive reasoning

The above discussion of weak and strong forms of abductive reasoning will be instrumental, in the rest of the paper, to understand and assess current methodological discussions within cognitive neuroscience. To this aim, it is crucial, as we argue, to introduce a novel distinction between what we shall call "weak" and "strong" reverse inference. Before coming to this, however, a closer look at how reverse inference is actually performed and analysed is in order.

3.1 Reverse inference in cognitive neuroscience

In cognitive neuroscience, neuroimaging techniques like fMRI play a crucial role. Roughly, fMRI allows researchers to find systematic correlations between cognitive processes plausibly engaged in experimental tasks and the increased activation, as measured by the BOLD activity in the relevant areas, of specific brain structures. As an example, participants in a typical fMRI experiment may be given tasks eliciting mental imagery or fear that, in turn, would be associated with increased activation of structures like the human precuneus or the amygdala, respectively (see Poldrack & Yarkoni, 2016). In this context, two different patterns of reasoning can be usefully distinguished.

Forward inference generally refers to reasoning from the (putative) engagement of a given cognitive process (e.g., fear) to the expected increased activation of a given structure of the brain (e.g., the amygdala).[3] In slightly more formal terms, forward inference concerns the probability $p(ACT|COG)$ of activation of some neural structure given the engagement of a given cognitive process. Such probabilities can be extracted by traditional meta-analysis of neuroimaging data, which highlight the regions of the brain that are more consistently associated with different cognitive processes. For example, suppose that the vast majority of studies involving a task (e.g., reading concrete words) that provoke an intense experience of mental imagery have found increased activation in the precuneus. From this, one may reasonably expect that, in a new task involving mental imagery, the increased activation of the precuneus will also be observed.

The second reasoning pattern is called in the neuroscientific literature *reverse inference*, since it is in a sense the inverse of forward inference. For example, from the data just considered above, we might be tempted to conclude that precuneus activity is a good *marker*, or predictor, of mental imagery. Thus, when we observe precuneus

[3] See Henson (2006) for a partly different characterization of forward inference. For a philosophical discussion of forward (and reverse) in reverse in correlation with neuropsychological data, see Machery (2012).

activity in a new data set, we might be inclined to conclude the engagement of mental imagery processes. Note that reverse inference involves the (inverse) probability $p(COG|ACT)$ that a given cognitive process is engaged when increased BOLD activity in a certain brain region is observed.

Although clearly related to each other, forward and reverse inference play crucially different roles, and their methodological status is also significantly different. As Poldrack (2011) notes, forward inference is «[t]he classic strategy employed by neuroimaging researchers», constituting «the basis for a large body of knowledge that has derived from neuroimaging research» (p. 692). In comparison, and despite its widespread use in many experimental studies using fMRI, reverse inference is much more problematic, as emphasised again by Poldrack in his seminal 2006 paper.[4] There are at least two reasons for this. The first is that reverse inference clearly instantiates a case of abductive, and hence non-deductive, reasoning, thus inheriting all problems and weaknesses of this kind of inference. Indeed, we can use the Peircean schema to rephrase reverse inference as follows:

> Activation pattern ACT is observed;
> But if process COG were engaged, ACT would be a matter of course,
> Hence, there is reason to suspect that COG is engaged.

Of course, as a form of abductive reasoning, reverse inference «is not deductively valid, but rather reflects the logical fallacy of affirming the consequent» (Poldrack, 2006, p. 60). The second reason why reverse inference is highly controversial is that, even when a strong correlation between a given cognitive process and some brain area is found via forward inference, this doesn't license, in general, comparatively strong reverse inferences. As noted by Yarkoni (2015a), the main epistemic weakness of this latter kind of inference becomes clear by construing reverse inference as a probabilistic argument:

> […] by observing that the probability of a particular pattern of brain activity conditional on a given mental state is not the same thing as the probability of a particular mental state conditional on a given pattern of observed brain activity […]. For example, if I know that doing a difficult working memory task produces activation in the dorsolateral prefrontal cortex (DLPFC) 80% of the time, I am not entitled to conclude that observing DLPFC activation in someone's brain implies an 80% chance that that person is doing a working memory task. To see why, imagine that a lot of other cognitive tasks—say, those that draw on recognition memory, emotion recognition, pain processing, etc.—also happen to produce DLPFC activation around 80% of the time. Then we would be justified in saying that all of these processes consistently produce DLPFC activity, but we would

[4] Indeed, this does not mean forward inference is immune to epistemic risks, as observed by Poldrack & Yarkoni (2016, pp. 589–590). In neuroimaging experiments, the subtraction method is generally used to identify which brain regions are activated by specific cognitive function. This consists in using carefully designed experimental conditions that are supposed to differ only with respect to one process of interest. The subtraction method is problematic because it relies on what has been called the "assumption of pure insertion", which has been subject to intense criticism in neuroscience (see Poldrack & Yarkoni 2016 for discussion).

have no basis for saying that DLPFC activation is specific, or even preferential, for any one of these processes.

To make this problem precise, Poldrack (2006) proposes a Bayesian reconstruction of reverse inference along the following lines:

$$p(COG|ACT) = \frac{p(ACT|COG) \cdot p(COG)}{p(ACT)}$$

Here, the posterior probability of the engagement of process COG given the activation of area ACT is computed, through Bayes theorem, in terms of the likelihood of COG (i.e., the probability of ACT given COG) and its prior probability given the task at issue. As usual, we can rewrite the denominator of the above formula as follows:

$$p(COG|ACT) = \frac{p(ACT|COG) \cdot p(COG)}{p(ACT|COG) \cdot p(COG) + p(ACT|not - COG)p(not - COG)}$$

This rewriting makes clear that the posterior probability $p(COG|ACT)$ crucially depends, as emphasised by Poldrack (2006), on the *selectivity* of the neural response. In other words, it depends on how likely the activation of the neural region in question is both in the presence and in the absence of the relevant cognitive process—i.e., both given *COG* and *not-COG*. According to Poldrack, the selectivity of the neural response is the greatest determinant of reverse inference: «[i]f a region is activated by a large number of cognitive processes, then activation in that region provides relatively weak evidence of the engagement of the cognitive process; conversely, if the region is activated relatively selectively by the specific process of interest, then one can infer with substantial confidence that the process is engaged given activation in the region» (2006, p. 32).

3.2 Neuroinformatics at the rescue

As the above quotation makes clear, reverse inference has been discussed from the very beginning in the context of justification rather than in the context of discovery. The main questions under investigation have been: How justified are reverse inferences? Can we systematically rely on reverse inferences in neuroimaging research? How can reverse inference be improved? These questions gave rise to an intense theoretical debate in philosophy and cognitive science (for a review, see Nathan & Del Pinal, 2017).[5] While the methodological debate developed, working neuroscientists devised more direct strategies to improve the reliability of reverse inferences in their daily

[5] Contributors to the debate have proposed quite different approaches to the issue of how to improve strong reverse inference. One sees the main problem in the fact that our *cognitive ontology*, namely our traditional taxonomy of mental functions and tasks, is outdated and intrinsically defective. The low selectivity of many brain regions might improve when cognitive functions are characterized at a higher level of abstraction (Price & Friston, 2005), or in more precise terms (Poldrack & Yarkoni, 2016). This approach, sometimes labelled "cognitive ontology revision" (Anderson 2015), has motivated the emergence of several computational approaches to mental functions taxonomies, such as the *Cognitive Atlas* (Poldrack et al., 2011), with the aim of systematizing and improving our ontology of mental concepts and tasks. A second approach tends to question the Bayesian reconstruction originally proposed by Poldrack (2006). Machery (2014), for

routine. To this purpose, they followed a main strategy that can be again traced back to Poldrack (2006). As seen above, the crucial problem of reverse inference is to assess the selectivity of brain regions, which is generally estimated on a purely informal basis (i.e., by means of manual search and qualitative reasoning on previous fMRI literature). In his paper (2006), Poldrack proposed to address this issue by using one of the several databases of fMRI results available on the Internet, such as BrainMap.[6] Poldrack's proposal promoted a big expansion in the use of databases of fMRI data and machine-learning techniques to formally quantify the strength of reverse inference, which has later become part of so-called cognitive neuroinformatics (see Poldrack & Yarkoni, 2016).

One of the first advancements in this field was the introduction of *NeuroSynth* (Yarkoni et al., 2011), an online platform that allows the synthesis of big datasets of neuroimaging results using an almost completely automatised method of search. Differently from BrainMap or similar databases, NeuroSynth exploits relatively simple text mining algorithms to automatically extract from the published articles two pieces of information: figures containing brain activations maps and specific terms of interest used at high frequency (more than 1 in 1000 words) in the text and referring to brain regions (e.g., *prefrontal cortex*), mental functions (e.g., *working memory*), and experimental tasks (e.g., *delayed response task*). This approach produces an extensive database of term-to-coordinates mappings, which is currently covering results from more than 14.000 studies. In line with a policy of open science and data sharing, the NeuroSynth database is made freely available through a web-based portal (www.neurosynth.org).

Using the web-based portal, it is possible to perform automated meta-analysis (*forward inference*) of hundreds of individual psychological concepts—such as *vision*, *audition*, *working memory*, *pain*, and so on—or psychological "topics", that is, clusters of semantically-related terms. Critically, the system can compute, for any given voxel V in the brain and any given term T in the database, the probability that V was reported as activated in a paper conditional on its mentioning or not T. By using T as a proxy for the engagement of the corresponding process *COG*, such probabilities

Footnote 5 continued
instance, argues that RI should be reformulated in purely "likelihoodist" terms, thus avoiding the tricky issue of assessing the prior probability of the hypotheses under examination. Others have proposed to conditionalize all probabilities in the Bayesian reconstruction of RI on the specific task used in the study (Del Pinal & Nathan, 2013; Hutzler 2014). A third proposal suggests that reverse inference may be improved, and the selectivity issue mitigated, by shifting the focus of the analysis from isolated brain regions to entire networks of regions (Glymour & Hanson, 2016; Klein, 2012). Finally, the use of multivariate neuroimaging techniques, such as *multivoxel pattern analysis* (MVPA), has been suggested as a fourth strategy to improve reverse inference, in line with the idea that inferences based on "pattern-decoding" can overcome the problems of more "local" ones (Nathan & Dal Pinal, 2017).

[6] In his paper (2006), Poldrack proposes to address this issue by using one of the several databases of neuroimaging results available on the Internet, i.e., BrainMap (www.brainmap.org), which at that time (Sept. 2005) contained data from 3222 experimental comparisons in 749 published papers. Looking at pairs of experimental comparisons and coordinates of activations included in this database, Poldrack manually calculated the probability of the engagement of language function conditional to the activation of the "Broca's area" (BA 44) using Bayes theorem. He later compared the posterior probability thus obtained (0.65) to the prior probability of language processes being engaged in a task, conventionally fixed at 0.5, and finally calculated the relative Bayes Factor (2.3) as a proxy of the strength of the reverse inference, resulting in a «positive but relatively weak increase of confidence» (p. 62) in the conclusion.

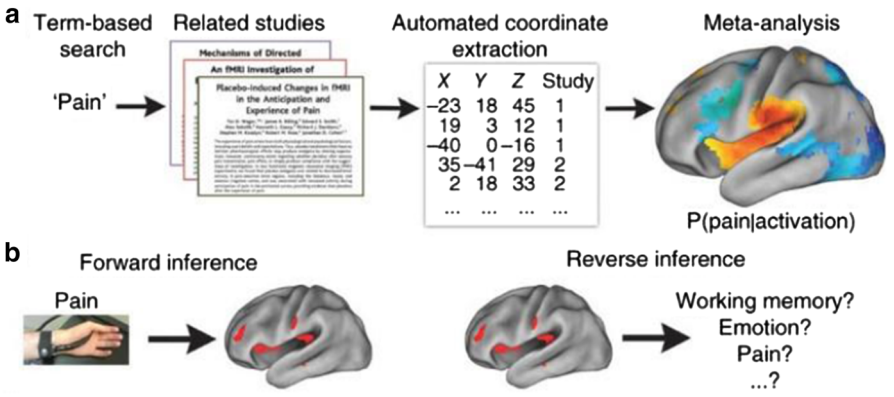

Fig. 1 A schematic representation of NeuroSynth's main functions. Modified with permissions from Yarkoni et al. (2011)

provide an estimate of the likelihoods $p(ACT|COG)$ and $p(ACT|notCOG)$. Then, at least in the intentions of its developers, NeuroSynth should apply Bayes theorem to compute the final posterior probability $p(COG|ACT)$, by assuming a default uniform prior (0.5) for $p(COG)$. These automated computations allow one to rank all relevant cognitive processes (i.e., psychological terms like *working memory*) by their posterior probability relative to any given coordinate of brain activation. Such "reverse inference maps"[7] allow, in turn, researchers to perform so-called *quantitative* reverse inferences (Yarkoni et al., 2011), where the strength of reverse inference is measured in terms of the automatically computed posterior probabilities. In the next section, we shall see how such a system can be used to make very strong claims about brain functioning (Fig. 1).

3.3 Reverse inference as abductive reasoning

What is critical for our purposes is that, despite Poldrack's initial remarks, these recent developments completely neglected the claim that reverse inference is a form of abductive reasoning. As a consequence, no attempt was made to apply the conceptual instruments developed in the philosophical literature to the case of cognitive neuroscience: most notably, the distinction between strong and weak abduction has been fully disregarded. In our opinion, this oversight has at least two critical consequences for the debate on reverse inference as it currently stands, which we shall analyse in detail in the next two sessions of this paper.

The first consequence is that, in general, scholars that have defended (or criticised) "strong reverse inference" as a form of justificatory abductive reasoning, have not properly appreciated its logical form. In its strong interpretation, reverse inference should be formulated as a rule of acceptance, that gives reasons to tentatively accept its conclusion as the "best" explanatory hypothesis among the available ones. This

[7] Note that "reverse inference maps" have been recently renamed "association tests" on the web-based NeuroSynth platform (https://neurosynth.org/faq/).

formulation of strong reverse inference as a form of IBE is hardly encountered in current debates; as we shall claim in Sect. 3, however, it is very useful to assess some of the most representative uses of fMRI data found in the neuroimaging literature, as well as the discussion that these representative uses have generated. Construing reverse inference as a form of IBE may help in evaluating the different comparative criteria for assessing competing cognitive explanations of activation patterns, a crucial issue that attracts the attention of neuroscientists but that still needs much philosophical and methodological work.

The second—and maybe more critical—consequence of overlooking the abductive nature of reverse inference is that what we called "weak reverse inference" has been virtually ignored in both philosophical and scientific discussion. This is problematic because, as we shall see in detail in Sect. 4, taking into account the strategic or discovery function of reverse inference is crucial to make sense of current neuroscientific practice. In many cases, reverse inference is indeed employed as a search strategy that tells us which hypotheses about the cognitive interpretations of a given brain activation we should set out for further inquiry and/or as a tool for making new hypotheses and assist discovery. Thus, exploring the nature and limits of weak reverse inference remains an important open task. Interestingly, the idea that reverse inference could be interpreted as a weak form of abductive reasoning was somehow foreshadowed in his 2006 paper by Poldrack himself, who noted that, «[v]iewed as a means to generate novel hypotheses, […] reverse inference can be a very useful strategy, especially if it is based on real data […], rather than an informal reading of the literature» (2011, p. 696). Despite occasional claims like this one, however, the strategic function of reverse inference has never been explored in detail so far. In the following, we shall fill this gap and offer the first comprehensive discussion of both weak and strong forms of reverse inference.

4 Strong reverse inference

As anticipated, the theoretical debate on reverse inference just surveyed has mainly focused on the justificatory role of reverse inference. Nevertheless, the nature of strong reverse inference as IBE has never been fully appreciated. To illustrate this point, we shall discuss first a representative case, which has generated an intense debate in cognitive neuroscience but has been completely ignored in the philosophical literature. Then, we will draw some lessons for the debate on reverse inference.

4.1 Strong reverse inference as a form of IBE

In a 2015 paper, neuroscientists Matthew Lieberman and Naomi Eisenberger used the NeuroSynth database to claim that the dorsal anterior cingulate cortex (dACC), a brain region that has been associated with several different cognitive functions, is actually selective for pain:

> Using Neurosynth, an automated brainmapping database, we performed quantitative reverse inference analyses to explore the best general psychological

> account of the dACC function P (Ψ process | dACC activity). Results clearly indicated that the best psychological description of dACC function was related to pain processing—not executive, conflict, or salience processing (2015, p. 15250).

Lieberman and Eisenberger used NeuroSynth to study the correlation between the activation of the dACC and four cognitive processes (including *pain*), known to be associated with the dACC. Their conclusion was that «whereas psychological processes and tasks related to pain, executive processes, conflict, and salience all reliably activate the dACC, the only psychological phenomenon that can be reliably inferred given the presence of dACC activity is pain» (2015, p. 15,522). More precisely:

> [a]lthough forward inference analyses reproduced the findings that many processes activate the dACC, reverse inference analyses demonstrated that the dACC is selective for pain and that *pain*-related terms were the single best reverse inference for this region (p. 15250).

Remarkably, their study relied on NeuroSynth as the only source of neuroscientific evidence. In particular, as evidence for the above claim, Lieberman and Eisenberger presented a comparison among the reverse inferences concerning the four different processes considered, and measured their relative strength relying on the statistics (i.e., posterior probabilities and associated Z-scores) for the four terms of interest across eight foci in the dACC (Fig. 2). Statistical analyses revealed that the Z-scores for *pain* were significantly greater than the Z-scores for each other terms of interest across all foci. Further analyses revealed that *pain*-related Z-scores across all foci of the dACC were greater than those related to each of the other terms in the NeuroSynth database (> 3000).

For our purposes, Lieberman and Eisenberger reasoning constitutes a spectacular example of what we call strong reverse inference*:* that is, a reverse inference whose conclusion is presented as strongly justified, or even as true, given its being the "best" of a series of alternative candidate hypotheses. This reasoning instantiates the "textbook version" of strong abduction (IBE) as presented by Douven (see Sect. 2), where the dACC activity represents the evidence E that should be explained, while the alternative psychological processes associated with the dACC activity represent the different candidate explanations $H_1,\ldots H_n$ of E. As observed by Wager (2015), Lieberman and Eisenberger's statement appears very strong and provocative, «as it implies that we can use these results to infer that dACC activity implies the presence of pain. After all, if the best "label" is pain, it seems like a reasonable inference».

The publication of Lieberman and Eisenberger's paper immediately triggered a hot debate, which has lasted a couple of years, with highly critical blog posts by), the creator of NeuroSynth, Alex Shackman (2015), Tor Wager (2015), and Lieberman himself (2015, 2016). After these informal exchanges, several influential names in the neuroimaging community published a commentary in PNAS (Wager et al., 2016) followed by a reply by Lieberman et al. (2016). In our opinion, this discussion is particularly representative of the kinds of difficulties that strong reverse inference, be it NeuroSynth-based or not, might encounter. Critically, many critiques advanced against the paper follow closely well-known philosophical objections to IBE in general.

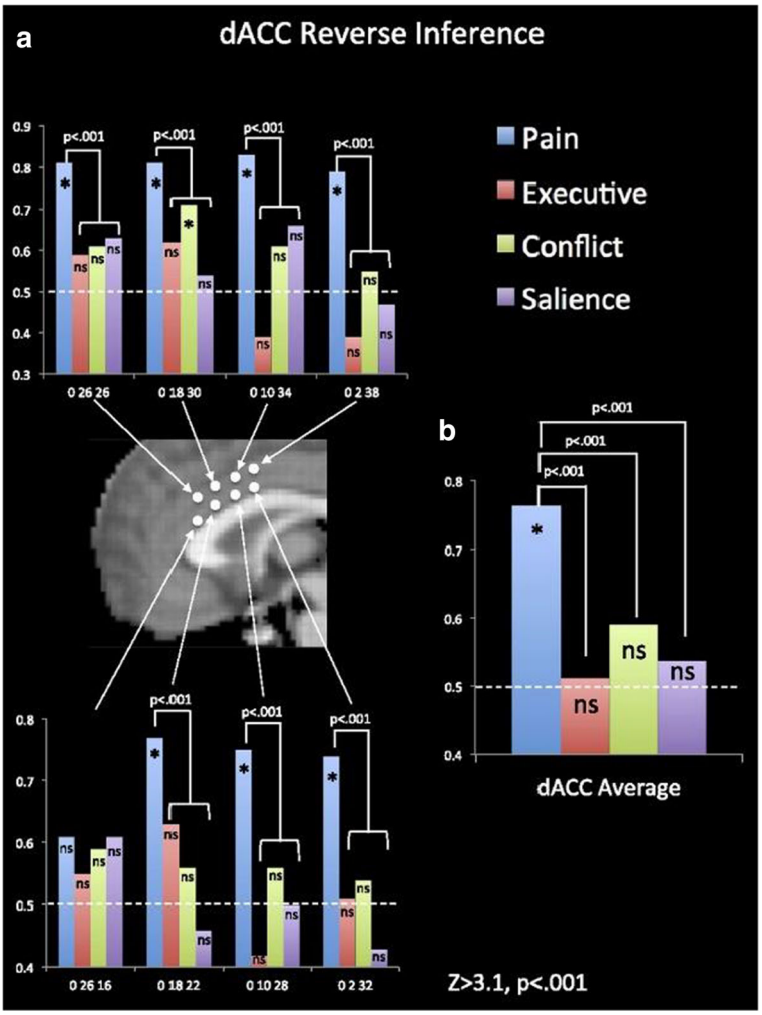

Fig. 2 Comparison of reverse inference effects throughout the dACC. Reproduced with permissions from Lieberman & Eisenberger (2016). See the text for further details

As an example, Yarkoni (2015a) focuses on the comparative analyses proposed by Lieberman and Eisenberger and criticises it as follows: «perhaps the most obvious problem is that it is largely based on comparison of *pain* with just three other groups of terms», thus possibly leading the authors to choose the hypothesis which was the best of a "bad lot". More importantly, Yarkoni (2015a) criticised the statistical method used by L&E to assess the relative strength of alternative reverse inferences. According to him, the authors correctly extracted from NeuroSynth the posterior probabilities associated with each term of interest, but then compared the associated Z-scores rather than the posterior probabilities themselves. This is a gross mistake because, from a Bayesian point of view, only posterior probabilities matter: «Z-score do not provide a measure of

strength of effect, they provide (at best) a measure of strength of evidence» (Yarkoni, 2015a).

Another critical worry is that terms in the NeuroSynth database have different base rates, since some cognitive processes are studied more frequently than others. For instance, *pain* occurs in only 3.5% of the neuroimaging studies, while *memory* occurs in 16% (Yarkoni et al. 2015b). As we have seen, NeuroSynth does explicitly account for frequency by setting the Bayesian prior for every term in the database to 0.5 (that is, the system uses *uniform priors* as default, rather than *empirical priors*). This default setting makes it possible to compare terms with significantly different frequencies in the database, but it does so by artificially masking the great variability in the base rates. In principle, it is possible to use the "core tools" of NeuroSynth to set the priors of the various terms to reflect the actual empirical frequency in the database. Nevertheless, we have no reason to assume that the empirical estimates of term frequency we can derive from NeuroSynth actually reflects the "real world" empirical priors ().

Based on this and similar criticisms, Yarkoni (2015a) concluded that NeuroSynth cannot be used to make strict comparisons between different candidate hypotheses and, more critically, to justify strong conclusions about the cognitive interpretation of a certain pattern of brain activity. To the extent that NeuroSynth is «one of the best tools we have at the moment» for justifying reverse inference, this seems to imply that strong reverse inferences (NeuroSynth-based or not) are never fully justified:

> NeuroSynth provides no license for saying much stronger things like "the dACC is selective for pain" or suggesting that one can make concrete reverse inferences about mental processes on the basis of observed patterns of brain activity. If the question we're asking is *what are we entitled to conclude* about *the presence of pain when we observed significant activation in the dACC in a particular study?*, the simple answer is: *almost nothing*.

4.2 The prospects of strong reverse inference

The discussion between Lieberman and Eisenberger, on the one hand, and Yarkoni, on the other, has been limited to these first semi-formal 2015–2016 exchanges. However, Yarkoni's conclusions above have helped to bolster a widespread scepticism over reverse inferences in the fMRI community. Also influenced by Poldrack's pioneering work, many cognitive neuroscientists now regard reverse inference in general as something that should be treated with much caution, or even as something that should be simply discarded (e.g., Anderson, 2010). Consequently, neuroimaging researchers applying reverse inferences are quickly criticised for committing the "fallacy of reverse inference", suggesting that this form of reasoning should always be avoided.

We believe that Yarkoni's pessimistic verdict is unwarranted. As we have argued, the Lieberman-Yarkoni exchange should be conceptualised as a discussion concerning the appropriate way of performing IBE in cognitive neuroscience; more precisely, as a discussion on how to compare the relative strength of alternative reverse inferences in the Bayesian framework underlying NeuroSynth (or similar database). If this is true, it is easy to understand that a proper debate on such an important issue is still

very much at a preliminary stage, and no explicit proposal has been made at the theoretical level. More critically, working cognitive neuroscientists are often not fully aware of the multiplicity of possible ways of theoretically addressing the problem; in actual practice, they tend to implicitly use different criteria for assessing competing explanatory hypotheses, which might easily lead to conflicting interpretations of the same experimental results.

For instance, as we have seen, Yarkoni suggests that the correct method for assessing the strength of reverse inference is to select the cognitive hypothesis with the highest posterior probability as computed from the NeuroSynth dataset. However, both Poldrack in his original theoretical analysis (2006) and other cognitive neuroscientists in actual experimental studies (e.g., Cauda et al., 2020), have proposed the Bayes factor as a criterion for selecting the best hypotheses in concrete reverse inferences. No explicit methodological discussion about the relation between these two measures can be found in the neuroscience literature. Interestingly, these two proposals in principle are not equivalent. As the discussion in philosophy of science has made clear (Sprenger & Hartman, 2019; Crupi, 2020), the former proposal amounts to construe evidential confirmation as "firmness" (how highly probable is COG given ACT); the latter instead defines confirmation as "increase in firmness" (how higher is $p(COG|ACT)$ than $p(COG)$). Critically, these two strategies of hypotheses assessment can lead to inconsistent results; in general, one can have high posterior probability without (incremental) confirmation, and vice versa. Thus, assessing the strength of a reverse inference using the posterior probability or the Bayes factor is *not* theoretically equivalent.

Moreover, posterior probabilities and the Bayes Factor are not the only options for assessing competing cognitive explanations of activation patterns. Indeed, philosophers have developed a number of formal measures for both the confirmation provided to competing hypotheses by a piece of evidence and for the explanatory power of these hypotheses (see, for instance, Schupabch & Sprenger, 2011). In principle, nothing prevents the implementation of such measures as comparative criteria within NeuroSynth or similar databases, although this possibility has never been explored so far. Importantly, however, neither Bayesian confirmation measures nor explanatory power measures are in general "ordinally equivalent". This means, roughly, that assessments of evidential favouring of one hypothesis over the other may be crucially sensitive to the choice of the underlying measure. As a consequence, assessing the strength of reverse inference will be always relative to the choice of the relevant measure: to our knowledge, however, no discussion of this point, which is both legitimate and urgent, appears in the literature.

Finally, even the Bayesian analysis which inspires the automated calculations performed by NeuroSynth is not without problems. One classical problem in Bayesian reasoning in general (Spenger & Hartman 2018), i.e., how to choose the prior probabilities to be used in Bayes' formula, is here especially serious. As we have seen, NeuroSynth assumes a flat distribution on the priors (in order to avoid selection bias from the current literature), but this assumption is arguably unsatisfactory without further justification. Although some suggestions for dealing with this problem have been recently advanced in the neuroscience literature (Costa et al., 2021), much work remains to be done. Moreover, this crucial issue adds to other known limitations of the

current version of NeuroSynth, such as the poor sensibility of the algorithm for terms and coordinates extraction (Yarkoni, 2015a).[8] Still, in our opinion, nothing prevents in principle that future developments of the database—like the NeuroSynth 2.0 system, which will be implemented in the more comprehensive platform NiMare[9]—might partially overcome these limitations. Furthermore, in recent years, it has been increasingly common to combine a NeuroSynth-based analysis with a discussion of independent evidence from patient data, TMS, and other techniques (e.g., Lieberman et al., 2019) and this may improve the prospects of such an approach to strong reverse inference.

To be sure, it is not our purpose in this context to discuss a method for making strong reverse inference in a correct way. Our only claim is that construing reverse inference as a form of IBE makes clear that strong reverse inference requires precise comparative criteria for assessing competing cognitive explanations of activation patterns. Thus, even if we believe that Yarkoni's pessimistic conclusion is not justified at the present state of knowledge, this crucial issue surely needs much philosophical and methodological work.

5 Weak reverse inference

As we argued in the foregoing section, the debate developed around Lieberman and Eisenberger's paper clearly shows that neuroscientists do rely, at least implicitly, on a strong understanding of reverse inference as IBE in their research. Indeed, we believe that the idea of strong reverse inference surely captures some representative case studies in the neuroimaging literature, even if not most of them. Focusing only on this strong reading of reverse inference (as it has been done until now), however, is arguably mistaken, for at least two reasons. First, as suggested above, this risks to fuel unwarranted scepticism toward reverse inference in general, as based on a (sound) criticism of strong reverse inference only. Second, we believe that many instances of reverse inference as actually performed in current neuroscientific research should not be construed as cases of IBE, but instead as attempts to discover new promising hypotheses to be assessed in further experimental studies. Such cases require a different kind of analysis, based on a weak notion of reverse inference, that has never been attempted so far. As a step forward in this direction, in the following we discuss in turn three representative case studies from recent neuroimaging research. As we argue, each of them clearly instantiates a case of weak abductive reasoning; together, they offer a fairly comprehensive view of the different functions that weak reverse inference may usefully perform in experimental studies. Interestingly, these functions reflect quite closely those attributed by philosophers to weak abduction in general (as discussed in Sect. 1). Such an analysis is thus interesting both on a theoretical and a practical level: indeed, neuroscientists rejecting the use of strong reverse inference may

[8] It is known that the automated lexical algorithms NeuroSynth is based on are not able to extract fine-grained information from texts (e.g., distinguishing different types of memory). Similarly, the algorithms extracting the coordinate of brain activations cannot make basic distinctions such as distinguishing between activations and deactivations (but see Yarkoni et al., 2011).

[9] See https://nimare.readthedocs.io/en/latest/.

still appreciate the role of weak reverse inference as an essential inferential strategy in cognitive neuroscience.

5.1 Weak reverse inference as selective abductive inference

As a first case study, consider Xenophontos and colleagues' examination of the effects of altered sex chromosome dosage (SCD) on brain functioning (Xenophontos et al., 2020). Altered SCD is a pathological dysfunction that characterises certain genetic disorders, for example, sex chromosome aneuploidy (SCA) syndromes. To identify the regions that are more affected by altered SCD, Xenophontos and colleagues tested 301 subjects affected by SCA and looked for regions with abnormal (i.e., increased or decreased) cortical thickness. Using structural MRI, the authors found that mounting SCD increased cortical thickness in the rostral frontal cortex (among other regions), bilaterally, and decreased cortical thickness in the lateral temporal cortex and in the temporal-parietal junction (among other regions). Finally, they relied on NeuroSynth to identify «cognitive and psychological terms that frequently cooccur in the literature with functional activations similar to the observed pattern of SCD effects on cortical morphology» (Xenophontos et al., 2020, p. 2219).

The authors compared the distribution of SCD effects on cortical thickness with the reverse inference maps (now "association maps") for each of 50 "topics" (clusters of semantically-related terms) included in the NeuroSynth database and then selected all the topics maps that showed a special correlation (|r| of 0.1 or greater) with the SCD maps. Results of this reverse inference analysis indicated that regions where, in SCA individuals' brains, cortical thickness increases as a function of mounting SCD are associated in the NeuroSynth database with emotion, pain, and inhibitory processing. Conversely, regions where cortical thickness decreases as a function of mounting SCD have been associated with visual, motor, arithmetic, and attentional processing. Since it is known that socioemotional and attentional processes are generally impaired in SCA individuals, the authors conclude that these «findings […] elucidate potential anatomical substrates for cognitive and behavioral alterations across SCA syndromes» (p. 2224).

Critically, these conclusions are not presented by the authors as strongly justified hypotheses, but rather as mere suggestions for further experimental research. As Xenophontos and colleagues recognise, «[m]ultimodal neuroimaging studies will be required to systematically assess the degree of overlap between structural and functional brain changes in SCA» (p. 2224). In other words, reverse inference is not used here to justify strong conclusions about brain functioning but only to generate a set of hypotheses (about the cognitive processes associated to the regions mostly lesioned in the SCA syndromes) that should be tested with other techniques to be empirically confirmed. According to some scholars, such as Tor Wager (2015), this is essentially the function of NeuroSynth: «NeuroSynth is a wonderful tool for hypothesis generation and for getting a rough idea of what a brain map related to a psychological topic might look like, but it was never intended to justify strong inferences about the psychological meaning of activation».

In the terminology introduced in Sect. 1, the study under examination is an example of *selective* abductive inference, where NeuroSynth is used as an artificial substitute for what Schurz (2017) calls probabilistic elimination techniques, suggesting a «short and promising (but not necessarily successful) path through the search space of possible explanatory hypothesis» (p. 153). Importantly, weak reverse inference is here used to suggest a restricted set of worthy candidate explanations for further experimental investigation, but not to generate a completely new hypothesis about the cognitive functions underpinned by the brain regions examined. In other words, it is a case of selective but not *creative* abductive reasoning in Magnani's sense.

Prima facie, one might doubt that NeuroSynth or similar meta-analytic tools can be of any help in generating truly creative weak abductive hypotheses. In fact, these systems strictly depend on already formulated cognitive hypotheses codified in published articles and previous neuroimaging literature. Consequently, the objection goes, they cannot foster new hypotheses about the cognitive functions associated to a given brain region. Nevertheless, we believe that this first impression is misleading. Indeed, NeuroSynth or similar tools might assist the discovery of associations between cognitive functions and brain structures that are present but still undetected in the neuroimaging data, as the next case-study shows.

5.2 Weak reverse inference as creative abductive inference (i)

As an example, consider Pauli and colleagues' investigation of the functional specialisation of the human striatum, a subcortical region that has been traditionally associated with emotion regulation and reward-related processes. In their study (Pauli et al., 2016), the authors first relied on NeuroSynth to identify distinct functional subregions in the striatum. Based on this analysis, they were able to identify five distinct striatal zones that exhibit discrete patterns of coactivation with distal cortical regions: ventral striatum, anterior and posterior caudate nuclei, anterior and posterior putamen.

Then, to identify which cognitive functions are more regularly associated with each striatal zone in the literature, Pauli and colleagues relied on NeuroSynth-based reverse inference. For each psychological term, the authors calculated the likelihood ratio as the «ratio of the number of studies reporting activation in a striatal sub-region when the term was vs. was not used in the article» (p. 1909). Results revealed some well-known associations between striatal regions and low-level cognitive functions, such as the association between ventral striatum and reward processing (i.e., with terms such as *rewards, losses,* or *craving*; p. 1909). However, the analysis also showed some associations that had not been highlighted in previous literature, thus «extend[ing] previous knowledge of the involvement of the striatum in reward-related decision-making tasks» (p. 1911):

> […] because we followed an unbiased, data-driven approach, we also identified associations between striatal activation and other psychological functions that have often been considered to be primarily cortical. In particular, cognitive functions, such as working memory and arithmetic, were associated with activation in the [posterior caudate], and social functions, such as language and empathy, were associated with activation in the [anterior putamen] (p. 1910).

According to the authors, the associations between higher-order psychological functions and striatal regions have gone undetected in previous literature because «the majority of studies investigating these psychological functions report activity preferentially in cortical areas, except for studies investigating reward-related and motor functions» (p. 1909). The specialisation of the posterior caudate for executive functions was particularly novel and unexpected, since these functions were «often considered the exclusive domain of the frontoparietal cortical circuits» (p. 1907).

For our purposes, this case illustrates well how data mining using meta-analytic tools such as NeuroSynth or BrainMap can be a tool of discovery. One might object that, in the example discussed above, the abductive reverse inference possesses a degree of novelty that is still too low to be considered a genuine case of creative abduction. In these kinds of abductions, according to Schurz, the «underlying abduction operation constructs something new, for example, a new theoretical model or a new theoretical concept» (2017, p. 158). In the context of cognitive neuroscience, a case of genuine creative abduction in this sense might be represented by an inference introducing a novel psychological concept, that is, a new entity in our scientific domain of cognitive functions, or *cognitive ontology* (see Sect. 2). This novel psychological concept might be a cognitive operation or sub-operations that has never been isolated before, such as a new type of memory. Alternatively, it might be a new general function subsuming most or all of the cognitive functions previously associated to a certain brain region—as in the case of the "sensorimotor integration" operation postulated by Price and Friston (2005) to explain why left posterior lateral fusiform is active in a vast range of behavioural domains. Nevertheless, the objection goes, NeuroSynth or similar tools cannot foster these classes of inferential operations. At most, these tools can sustain conceptual rearrangement but not concept creation in Peirce's terms (see Sect. 1). Again, we believe that this first impression is misleading.

5.3 Weak reverse inference as creative abductive inference (ii)

In a recent paper, Genon et al. (2017) investigated the connectivity patterns and the functional organisation of the left dorsal premotor cortex (PMd). As a result of their analysis, they identified five functionally distinct sub-regions of PMd: rostral, central, caudal, ventral, and rostro-ventral. Then, Genon and colleagues relied on the BrainMap database (Laird et al., 2005) to characterise the functional profile of each of the PMd sub-region. Relying on Bayes theorem for a quantitative reverse inference analysis, they calculated which were the most likely cognitive domains (i.e., "behavioural domain" in the Brain Map's terminology) and the most likely experimental tasks (i.e., "paradigm class") conditional on the activation in each sub-region of PMd. Reverse inference across behavioural domains and experimental paradigms revealed that PMd is a highly multifunctional region, with different classes of cognitive processes associated with distinct sub-regions of PMd (Fig. 3).

Critically, in order to explain the specific functional profile of a particular subregion of PMd, i.e., the rostro-ventral PMd, Genon and colleagues introduced a novel psychological entity that was never discussed before. In the BrainMap database, the «rostro-ventral PMd was associated with tasks related to explicit long-term memory,

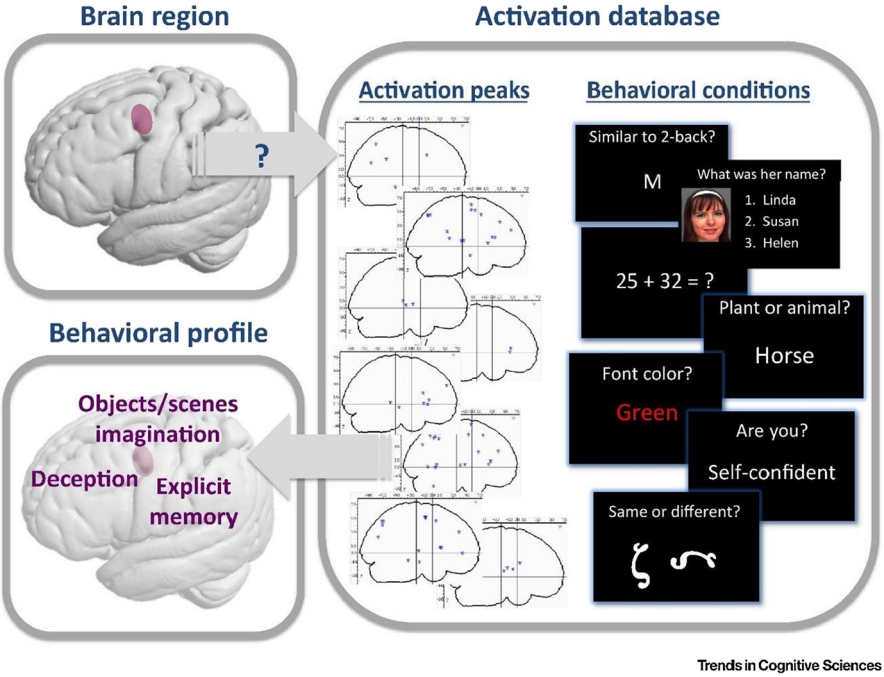

Fig. 3 A schematic representation of the quantitative reverse inference on the BrainMap database by Genon and colleagues (2017). Reproduced with permissions from Genon et al. (2018)

object or scene imagination, and deception paradigms» (p. 410). Based on such results, the authors speculated that such region underpins an abstract cognitive process, a «core computational function», which grounds all these behavioural associations but remains latent and is not directly observed. In a subsequent paper (2018), Genon and colleagues better characterised the abstract cognitive operation underpinned by the rostro-ventral PMd introducing the concept of "sequential processing":

> [w]e can speculate that this abstraction property reflects the use of sequential processing (spatial or temporal) in the PMd for various types of predictions beyond the current framework, in line with the Bayesian brain hypothesis (p. 357).

Interestingly, Genon and colleagues' reasoning fits quite well with what Schurz calls "hypothetical (common) cause abductions", a kind of abduction that postulates a new entity or property to explain a set of empirical phenomena that were previously considered as unrelated (in this case, the set of behavioural tasks associated to activation in the ventro-rostral PMd). According to Schurz (2017), «this is the most fundamental kind of conceptually creative abduction», which is driven by «the search of explanatory *unification*» (p. 162). As a consequence, the author's hypothesis has the potential to trigger a *cognitive ontology revision* (see Anderson, 2015), devising

an entirely new cognitive kind. In turn, suggesting a more risky but potentially fruitful line of research, their abductive inference complies with the Peircean *dictum* of preferring "uberty" over "security".

It is important to note that, as recognised by Genon and colleagues, these kinds of observations are only possible when neuroimaging researchers rely on quantified reverse inference with NeuroSynth, BrainMap, or similar meta-analytic tools, which allow for the integration activations across thousands of different tasks and behavioural domain:

> [a]s illustrated in [these examples], the patterns of associations across a wide range of tasks can foster new hypotheses, approximating as much as possible the core role of the region (and thus its operation-function), beyond the behavioural ontology of the original studies or the database (Genon et al., 2018, p. 357)

Genon and colleagues' reasoning is not an isolated case in the neuroimaging literature. For instance, different studies employing the NeuroSynth database have shown that the anterior insula is engaged in a wide range of fMRI tasks; on this basis, it has been suggested that this area supports a novel generic function, i.e., "task engagement maintenance" (see Poldrack, 2011). In these and other similar cases, we can construe reverse inference as a form of weak abducting reasoning that can suggest radically new hypothesis about the cognitive function associated to a given brain region. When successful, this kind of reasoning seems to support what Peirce called "concept creation", thus showing that NeuroSynth or similar tools can also be applied for performing weak, creative forms of reverse inference.

6 Conclusion

In this paper, we offered the first comprehensive discussion of reverse inference as a form of abductive reasoning in cognitive neuroscience. Relying on previous philosophical discussion, we first introduced a distinction between two forms of this inferential strategy, i.e., weak and strong reverse inference. Then, we argued that distinguishing between these two functions of reverse inference—i.e., a justificatory (strong) and a strategic (weak) one—is crucial both to make sense of current neuroscientific practice and for assessing the methodological debate on reverse inference in general. In support of this, we provided the first systematic exploration of both the justificatory and the strategic function of reverse inference.

The main results of our discussion can be summarised as follows. On the one hand, strong reverse inference as a form of IBE clearly plays a role in some of the boldest attempts of deriving conclusions about the engagement of cognitive processes based on fMRI data found in the neuroimaging literature. Accordingly, looking at this pattern of reasoning as a form of strong abduction can surely advance the discussion on the justificatory role of reverse inference, especially for what concerns precise comparative criteria for assessing competing cognitive explanations of activation patterns. At the present state of research, however, such a discussion is absent from the neuroscientific debate, and this has fuelled an undue pessimism on the reliability of reverse inference in general.

On the other hand, the weak function of reverse inference has been virtually ignored in both the philosophical and neuroscientific literature. Still, weak reverse inference is indeed performed in current neuroscientific research and, as our discussion reveals, instantiates most of the strategic functions that philosophers have traditionally assigned to abduction. In particular, we examined three case-studies, illustrating both the selective function of weak reverse inference (i.e., individuating a restricted set of plausible hypotheses worth of further empirical testing) and its creative function (i.e., suggesting a partially or radically new psychological interpretation of a given brain activation). Acknowledging the role of weak reverse inference in current research practice thus sheds new light on its methodological role and may mitigate the scepticism that presently surrounds reverse inference within the community.

Of course, the present paper has provided just the beginning of a more systematic exploration of the role of abductive reasoning in cognitive neuroscience; the spirit of the paper, hence, is programmatic. Indeed, we believe that the distinction between weak and strong reverse inference might shed new light on the debate on reverse inference as it currently stands, clarifying some important issues and even opening new directions for methodological reflection in the field. To this purpose, a more detailed study of both the justificatory and the strategic function of reverse inference, as well as of the role of NeuroSynth and similar tools in supporting such functions, is surely needed. With the present contribution, we hope we provided a general framework to rigorously address such problems in future research.

Acknowledgements Previous versions of this article have been presented at various conferences, such as the workshop on "Scientific Errors" (Castelveccana, Italy, 2021) and the symposium on "Reverse Inference: Philosophical and Neuroscientific Perspectives" (ESPP, Online, 2021). We thank the participants at these conferences for their thoughtful comments. We are particularly grateful to Luca Cecchetti, Davide Coraci, Enzo Crupi, Enzo Fano, Diego Marconi, Jan Sprenger, Marco Viola, and two anonymous reviewers for Synthese for very useful comments on this article and/or discussions on its contents.

Funding Gustavo Cevolani acknowledges financial support from the Italian Ministry of Education, Universities and Research (MIUR) through the grant n. 201743F9YE (PRIN 2017 project "From models to decisions").

References

Anderson, D. R. (1987). *Creativity and the philosophy of C.S. Peirce*. Springer.

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

Anderson, M. (2015). Minding the brain for a new taxonomy of the mind. *Philosophy Compass, 10*(1), 68–77.

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

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's Inference to the best explanation. *Synthese, 180*, 419–442.

Cauda, F., Nani, A., Liloia, D., Manuello, J., Premi, E., Duca, S., Fox, P. T., & Costa, T. (2020). Finding specificity in structural brain alterations through Bayesian reverse inference. *Human Brain Mapping, 41*(15), 4155–4172.

Costa, T., Manuello, J., Ferraro, M., Liloia, D., Nani, A., Fox, P. T., Lancaster, J., & Cauda, F. (2021). BACON: A tool for reverse inference in brain activation and alteration. *Human Brain Mapping, 42*(11), 3343–3351.

Crupi, V. (2020). Confirmation. *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/confirmation/>.

Del Pinal, G., & Nathan, M. J. (2013). There and up again: On the uses and misuses of neuroimaging in psychology. *Cognitive Neuropsychology, 30*(4), 233–252.

Douven, I. (2017), Abduction. *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/abduction/>.

Douven, I., & Schupbach, J. N. (2015). probabilistic alternatives to Bayesianism: The case of explanationism. *Cognition, 6*, 459.

Frankfurt, H. (1958). Peirce's notion of abduction. *Journal of Philosophy, 55*, 593–596.

Genon, S., Reid, A., Langner, R., Amunts, K., & Eickhoff, S. B. (2018). How to characterise the function of a brain region. *Trends in Cognitive Sciences, 22*(4), 350–364.

Genon, S., Reid, A., Li, H., Fan, L., Müller, V. I., Cieslik, E. C., Hoffstaedter, F., Langner, R., Grefkes, C., Laird, A. R., Fox, P. T., Jiang, T., Amunts, K., & Eickhoff, S. B. (2017). The heterogeneity of the left dorsal premotor cortex evidenced by multimodal connectivity-based parcellation and functional characterisation. *NeuroImage, 170*, 400–411.

Glymour, C., & Hanson, C. (2016). Reverse inference in neuropsychology. *The British Journal for the Philosophy of Science, 67*(4), 1139–1153.

Greene, J. D., Sommerville, B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An FMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Hanson, N. R. (1958). *Patterns of discovery*. Cambridge University Press.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review, 74*, 88–95.

Henson, R. (2006). Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Sciences, 10*(2), 64–69.

Hintikka, J. (1998). What is abduction? The fundamental problem of contemporary epistemology. *Transactions of the Charles S. Peirce Society, 34*(3), 503.

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

Klein, C. (2012). Cognitive ontology and region- versus network-oriented analyses. *Philosophy of Science, 79*(5), 952–960.

Laird, A. R., Lancaster, J. L., & Fox, P. T. (2005). BrainMap: The social evolution of a functional neuroimaging database. *Neuroinformatics, 3*, 65–78.

Lieberman, M. (2015). Comparing Pain, Cognitive, and Salience Accounts of dACC. A reply to Tal Yarkoni's blog on our paper. URL: https://www.psychologytoday.com/us/blog/social-brain-social-mind/201512/comparing-pain-cognitive-and-salience-accounts-dacc.

Lieberman, M. (2016). More Evidence for a Pain-Related Description of dACC. Selective voxels in the dACC tend to be selective for pain. https://www.psychologytoday.com/ca/blog/social-brain-social-mind/201601/more-evidence-pain-related-description-dacc.

Lieberman, M.D., Burns, S.M., Torre, J.B., Eisenberger, N.I. (2016). Reply to Wager et al.: Pain and the dACC: The importance of hit rate-adjusted effects and posterior probabilities with fair priors. *Proc Natl Acad Sci USA*, *113*(18), E2476–9.

Lieberman, M. D., & Eisenberger, N. I. (2015). The dACC is selective for pain. *Proceedings of the National Academy of Sciences, 112*(49), 15250–15255.

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience and Biobehavioral Reviews, 99*, 311–328.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). Routledge/Taylor and Francis Group.

Machery, E. (2012). Dissociations in neuropsychology and cognitive neuroscience. *Philosophy of Science*, *79*(4), 490–518.

Machery, E. (2014). In defense of reverse inference. *The British Journal for the Philosophy of Science, 65*(2), 251–267.

Magnani, L. (2001). *Abduction*. Springer.

Magnani, L. (2009). Creative Abduction and Hypothesis Withdrawal. In J. Meheus & T. Nickles (Eds.), *Models of Discovery and Creativity. Origins: Studies in the Sources of Scientific Creativity.* Springer.

Mcauliffe, W. (2015). How did Abduction Get Confused with Inference to the Best Explanation? *Transactions of the Charles S. Peirce Society, 51*(3), 300–319.
McKaugan, D. (2008). From ugly duckling to swan C.S. Peirce, abduction, and the pursuit of scientific theories. *Transactions of the Charles S. Peirce Society, 44*(3), 446–468.
Minnameier, G. (2004). Peirce-suit of truth—why inference to the best explanation and abduction ought not to be confused. *Erkenntnis, 60*, 75–105.
Nathan, M. J., & Del Pinal, G. (2017). The future of cognitive neuroscience? reverse inference in focus. *Philosophy Compass, 12*(7), e12427.
Niiniluoto, I. (1999). Defending abduction. *Philosophy of Science, 66*(3), 451.
Niiniluoto, I. (2011). Abduction, tomography, and other inverse problems. *Studies in History and Philosophy of Science Part A, 42*(1), 135–139.
Niiniluoto, I. (2018). *Truth-Seeking by Abduction*. Springer.
Norton, J. D. (2016). Inference to the best explanation: The general account. In J. Norton (Ed.), *The material theory of induction.* University of Calgary Press.
Paavola, S. (2004). Abduction as a logic and methodology of discovery: The importance of strategies. *Foundations of Science, 9*, 267–283.
Paavola, S. (2006). Hansonian and harmanian abduction as models of discovery. *International Studies in the Philosophy of Science*, *20*(1), 93–108.
Pauli, W. M., O'Reilly, R. C., Yarkoni, T., & Wager, T. D. (2016). Regional specialisation within the human striatum for diverse psychological functions. *Proceedings of the National Academy of Sciences of the United States of America, 113*(7), 1907–1912.
Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.
Poldrack, R. (2008). The role of FMRI in cognitive neuroscience: Where Do We Stand? *Current Opinion in Neurobiology, Cognitive Neuroscience, 18*(2), 223–227.
Poldrack, R. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron, 72*(5), 692–697.
Poldrack, R., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D. S., Sabb, F. W., & Bilder, R. M. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics, 5*, 17.
Poldrack, R., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: Informatics and the search for mental structure. *Annual Review of Psychology, 67*, 587–612.
Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology, 22*(3–4), 262–275.
Roche, W., & Sober, E. (2013). Explanatoriness Is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory. *Analysis, 73*(4), 659–668.
Salmon, W. (2001). Explanation and confirmation: A Bayesian critique of inference to the best explanation. In G. Hon & S. S. Rakover (Eds.), *Explanation: Theoretical approaches and applications.* Kluwer.
Schackman, A. (2015). The importance of respecting variation in cingulate anatomy: Comment on Lieberman & Eisenberger 2015 and Yarkoni. URL: https://shackmanlab.org/the-importance-of-respecting-variation-in-cingulate-anatomy-comment-on-lieberman-eisenberger-2015-and-yarkoni/
Schickore, J. (2018). Scientific Discovery. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/scientific-discovery/>.
Schupbach, J. N. (2014). Is the bad lot objection just misguided? *Erkenntnis, 79*(1), 55–64.
Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science, 78*(1), 105–127.
Schurz, G. (2008). Patterns of abduction. *Synthese, 164*, 201–234.
Schurz, G. (2017). Patterns of Abductive Inference. In L. Magnani & L. Bortolotti (Eds.), *Springer Handbook of Model-Based Science.* Springer.
Sprenger, J., & Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford University Press.
Van Frassen, B. (1989). *Laws and Symmetry*. Oxford University Press.
Wager, T. (2015). Pain in the ACC. A commentary on Lieberman and Eisenberger. URL: https://www.painresearchforum.org/news/blog/61907-pain-acc
Wager, T. D., Atlas, L. Y., Botvinick, M. M., Chang, L. J., Coghill, R. C., Davis, K. D., Iannetti, G. D., Poldrack, R. A., Shackman, A. J., & Yarkoni, T. (2016). Pain in the ACC? *Proceedings of the National Academy of Sciences, 113*(18), E2474–E2475.

Weiskopf, D. (2020). Data Mining the Brain to Decode the Mind. In F. Calzavarini, M. Viola, *Neural Mechanisms. New Challenges in the Philosophy of Neuroscience*. Springer.

Xenophontos, A., Seidlitz, J., Liu, S., Clasen, L. S., Blumenthal, J. D., Giedd, J. N., Alexander-Bloch, A., & Raznahan, A. (2020). Altered sex chromosome dosage induces coordinated shifts in cortical anatomy and anatomical covariance. *Cerebral Cortex, 30*(4), 2215–2228.

Yarkoni, T. (2015a). No, the dorsal anterior cingulate is not selective for pain: comment on Lieberman and Eisenberger. URL: https://www.talyarkoni.org/blog/2015a/12/05/no-the-dorsal-anterior-cingulate-is-not-selective-for-pain-comment-on-lieberman-and-eisenberger-2015a/

Yarkoni, T. (2015b). Still not selective: comment on comment on comment on Lieberman & Eisenberger. URL: https://www.talyarkoni.org/blog/2015b/12/14/still-not-selective-comment-on-comment-on-comment-on-lieberman-eisenberger-2015b/

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.