Meet me in the middle: brain-behavior mediation analysis for fMRI experiments

Jules Brochard[123], Jean Daunizeau[123]

[1] Université Pierre et Marie Curie, Paris, France

[2] Institut du Cerveau et de la Moelle épinière, Paris, France

[3] INSERM UMR S975

Address for correspondence:

Jean Daunizeau

Motivation, Brain and Behaviour Group

Brain and Spine Institute

47, bvd de l'Hopital, 75013, Paris, France.

Tel: +33 1 57 27 43 26

Mail: jean.daunizeau@gmail.com

Abstract:

Functional outcomes (e.g., subjective percepts, emotions, memory retrievals, decisions, etc...) are partly determined by external stimuli and/or cues. But they may also be strongly influenced by (trial-by-trial) uncontrolled variations in brain responses to incoming information. In turn, this variability provides information regarding how stimuli and/or cues are processed by the brain to shape behavioral responses. This can be exploited by brain-behavior mediation analysis to make specific claims regarding the contribution of brain regions to functionally-relevant input-output transformations. In this work, we address four challenges of this type of approach, when applied in the context of mass-univariate fMRI data analysis: (i) we quantify the specificity and sensitivity profiles of different variants of mediation statistical tests, (ii) we evaluate their robustness to hemodynamic and other confounds, (iii) we identify the sorts of brain mediators that one can expect to detect, and (iv) we disclose possible interpretational issues and address them using complementary information-theoretic approaches. *En passant*, we propose a computationally efficient algorithmic implementation of the approach that is amenable to whole-brain exploratory analysis. We also demonstrate the strengths and weaknesses of brain-behavior mediation analysis in the context of an fMRI study of decision under risk. Finally, we discuss the limitations and possible extensions of the approach.

Introduction

Functional outcomes (e.g., subjective percepts, emotions, memory retrievals, decisions, etc...) are partly determined by external stimuli and/or contextual cues. But they may also be strongly influenced by irreducible variability in brain responses to incoming information (Ferster, 1996; Shadlen and Newsome, 1994). In particular, neural noise may be a critical determinant of illusory percepts, aberrant emotions, erroneous memory retrievals, biased decisions, etc... (Bays, 2014; Faisal et al., 2008; Hong and Rebec, 2012). For most existing statistical data analyses of neurophysiological data, neural noise is typically treated as a statistical nuisance, since it compromises the identification of relationships between measured brain activity and experimental variables (Doi and Lewicki, 2011; Naselaris et al., 2011). This perspective is unfortunate however, since neural noise can provide complementary information regarding how incoming information is processed and/or distorted by the brain to yield functional outcomes (Dinstein et al., 2015; McDonnell and Ward, 2011; Stein et al., 2005). The critical point is that a brain system may encode functionally-relevant information that is not *used* by the brain when producing a functional outcome. This has been repeatedly demonstrated in neurological patients who do not exhibit significant behavioral impairments despite being lesioned in brain regions that are known to encode behaviorally-relevant information(Aerts et al., 2016; Alstott et al., 2009). But what if one can show that neural noise contributes to -otherwise unexplained- behavioral variability? This is the essence of *brain-behavior mediation analysis*, which aims at detecting neural systems that both respond to behaviorally-relevant cues or stimuli and eventually impact overt behavior (MacKinnon et al., 2007).

Recall that any cognitive function can be seen as some form of -potentially complex, context-dependent, redundant, partially unconscious, etc- neural transformation of relevant stimuli into adaptive behavioural outcomes (Robbins, 2011). By adaptive, we simply mean that cognitive functions serve a specific purpose, which can be abstracted and put to a (behavioural) test. At the limit, one could argue that understanding cognitive functions reduces to assessing input-output relationships, where inputs are experimentally controlled stimuli and/or task instructions, and outputs are overt behavioural outcomes. In this view, neuroimaging in healthy subjects should serve to identify how brain networks contribute to the input-output transformation (Palestro et al., 2018; Rigoux and Daunizeau, 2015; Turner et al., 2019). A reasonable strategy here is to identify intermediary neural states that *mediate* the impact of incoming information onto overt behavior and/or subjective reports. In its simplest form, brain-behavior mediation analysis reduces to a twofold regression analysis that aims at detecting uncontrolled variability in brain responses that significantly improves behavioral predictability. The ensuing statistical tests typically reason as follows: if region M responds to experimental factor X, and explains behaviour Y above and beyond the effect of X, then M mediates the effect of X onto Y. For example, brain-behavior mediation analysis was used to identify the prefrontal and/or subcortical systems that mediate successful emotional regulation (Wager et al., 2008), threat response (Wager et al., 2009a, 2009b) or risk avoidance (Yamamoto et al., 2015). More recently, the anterior cingulate cortex, the anterior insula, the thalamus and some brain stem nuclei were shown to mediate various aspects of pain perception (Atlas et al., 2010, 2014; Geuter et al., 2018; Koban et al., 2017, 2019; Woo et al., 2015). Most of these studies were performed using the multilevel mediation/moderation or M3 toolbox (Wager, 2008), which was first derived for probing effective connectivity from fMRI signals. Since then, a few multivariate extensions of brain-behavior mediation analysis were proposed, aiming at improving either spatial or temporal resolution (Chén et al., 2018; Lindquist, 2012; Zhao and Luo, 2017). But these approaches neither lay out nor address the specific methodological and interpretational challenges posed by brain-behavior analysis, when applied to typical fMRI experiments. In our view, progress in brain-behavior mediation analysis requires answering at least four important (and related) questions:

- (Q1) Which test statistics should be used? Not only should the test statistics be valid (i.e. yield controlled false positive rate), but they also should be maximally powerful. The latter is a pressing issue because fMRI induces a massive multiple comparison problem, which can only be solved by using more stringent significance thresholds (Lindquist and Mejia, 2015; Worsley and Friston, 1995). We will summarize and compare the statistical properties of the most established test statistics of mediation analysis.
- (Q2) How robust is brain-behavior mediation analysis to assumptions regarding the hemodynamic response function (HRF) and other confounds? Recall that virtually all forms of fMRI time-series analyses rely on HRF models to assess effects of interest (Deshpande et al., 2010; Gitelman et al., 2003; Liao et al., 2002; Pedregosa et al., 2015). Although brain-behavior mediation analysis involves similar assumptions, different modelling strategies may be employed that yield distinct bias-variance tradeoffs. We will compare the statistical properties of these candidate approaches in the presence of deviations to modelling assumptions.
- (Q3) What sort of brain mediators can we expect to detect? Consider the bottom-up chain of neural information processing stages that eventually yield behavioral outcomes (from low-level sensory processing to high-level cognitive treatment of stimuli and/or cues). It turns out that these stages do not have the same chance of being detected. As we will see, this is a corollary consequence of the nontrivial

(and yet undisclosed) impact of neural noise onto the statistical properties of mediation analysis.

- (Q4) Does mediation analysis induce potential interpretational issues? As we will see, some interpretational issues are specific to the chosen statistical testing approach, but others are generic to any brain-behavior mediation analysis. In particular, significant mediated effects are compatible with two distinct scenarios regarding the causal relationship between brain activity and behavioral responses. We discuss the importance of this and related issues and identify ways to address them.

In this work, we address these four questions from a user-oriented statistical perspective. Our aim here is to set a methodological standard for brain-behavior mediation analysis. The Methods section serves as the statistical and conceptual basis for addressing the four questions (Q1-Q4) above. It starts with a description of the brain-behavior mediation model and its associated null-hypothesis testing alternatives. Specific issues that arise in the context of typical fMRI experiments (factorial designs and condition contrasts, group-level random effects analysis, etc) are shortly discussed. We then consider the critical role of neural noise in brain-behavior mediation analyses, and present alternative solutions to the issue of HRF deconvolution. We close this section with a note on causality and its accompanying interpretational issue. We address the latter using a complementary information-theoretic approach (so-called *I/O test*). *En passant*, we show how to exploit the underlying mathematical degeneracy to drastically reduce the computational cost of whole-brain mediation analysis. In the Results section, we use numerical Monte-Carlo simulations to answer questions Q1-Q4. We compare the specificity and sensitivity of candidate mediation tests, as a function of neural noise, and in the presence of hemodynamic confounds. We also evaluate the utility and robustness of our I/O test. We then strengthen our *in-sillico* conclusions with an application to an experimental fMRI dataset acquired when people make decisions under risk. We exemplify the use of brain-behavior mediation analysis to ask questions regarding intra- and between-subjects variations in behavioral responses and attitudes. Finally, we discuss our results in the light of the existing literature and highlight potential weaknesses and perspectives (Discussion section).

Methods

In what follows, we will consider behavioral paradigms akin to decision tasks, whereby subjects need to process some (experimentally-controlled) information X to provide a (measured) behavioral response Y. Brain-behavior mediation analysis then aims at identifying whether some (anatomically-specific) feature of their observed brain activity M mediates the effect of X onto Y. In our example fMRI application (see Results section), we will focus on a value-based decision making task, whereby participants have to accept or reject (response Y) a risky gamble composed of a 50% chance of winning a gain G and a 50% chance of loosing L (input information X). But more generally, X is an experimental manipulation of some sort, M is a measure of neural activity at the time of processing the stimulus, and Y is some overt expression of the stimulus-induced covert mental state of interest.

The brain-behavior mediation model

Let n be the number of trials in a typical experimental session. Let X, M and Y be $n\times 1$ column vectors encoding the trial-by-trial experimental manipulation, the brain's response to the experimental manipulation (e.g., the magnitude of the fMRI BOLD response to the stimulus at each trial, in some voxel or region of interest) and the behavioral response to the experimental manipulation, respectively. For the sake of mathematical simplicity, and without loss of generality, we will assume that X, M and Y have all been z-scored.

From the perspective of identifying the determinants of behavior, one may first ask whether X has an effect on Y or not. In its simplest mathematical form, this question reduces to considering the following simple linear regression model:

$$Y = Xc + \varepsilon_Y^0 \tag{1}$$

where c is an unknown regression coefficient that measures the strength of the statistical relationship between the independent (X) and dependent (Y) variables, and ε_Y^0 are model residuals. On would then simply test for the statistical significance of c, under some assumptions regarding the distribution of model residuals ε_Y^0.

Now, one may also ask whether M mediates the effect of X onto Y. In its simplest mathematical form, this question relies on the following pair of linear regression models:

$$\begin{cases} M = Xa + \varepsilon_M^0 \\ Y = Mb + Xc' + \varepsilon_Y \end{cases} \tag{2}$$

where the first equation expresses the fact that M responds to X (with some unknown susceptibility a), and the second equation states that Y depends upon both M (with some unknown susceptibility b) and X (with some unknown susceptibility c'). On may think of residuals ε_M^0 in terms of some form of *neural noise*, because they capture trial-by-trial variations in M that are independent of X. As we will see, they play a pivotal role in brain-behavior mediation analysis.

Although simple, Equation 2 does not explicitly quantify the size of a mediated effect. But this can be done by noting that Equation 2 can be rewritten as follows:

$$\begin{aligned} Y &= Xab + \varepsilon_M^0 b + Xc' + \varepsilon_Y \\ &= X\left(ab + c'\right) + \varepsilon_M^0 b + \varepsilon_Y \end{aligned} \qquad (3)$$

where M has simply been replaced by its expression from Equation 2. Equation 3 is helpful in realizing that the *total* effect of X onto Y is partitioned into a *direct* effect (whose size is c') and an *indirect* effect (whose size is ab). This distinction is important because the latter is the effect of X onto Y that is mediated by M. This is why established mediation tests rely on assessing the statistical significance of the indirect effect (MacKinnon et al., 2007). Note that so-called *full mediation* occurs when $c' = 0$ (no direct path), and one speaks of *partial mediation* whenever $c' \neq 0$.

Importantly, when we perform mass-univariate mediation analysis, we effectively consider each voxel or region of interest in isolation, and ask whether the local indirect effect is statistically significant. If mediation tests are repeated over voxels, then they form a statistical mediation *map*, which can localize which brain structure(s) mediate(s) the effect of X onto Y. In this context, Equations 1-3 have two interesting implications, which we will highlight now.

To begin with, recall that the incoming information X is processed by a distributed brain system, whose elements (sampled across large voxel sets) concurrently contribute to the behavioural response Y. The structure of this distributed brain system is likely to involve multiple processing pathways that work both in series and in parallel, as in Figure 1 below.

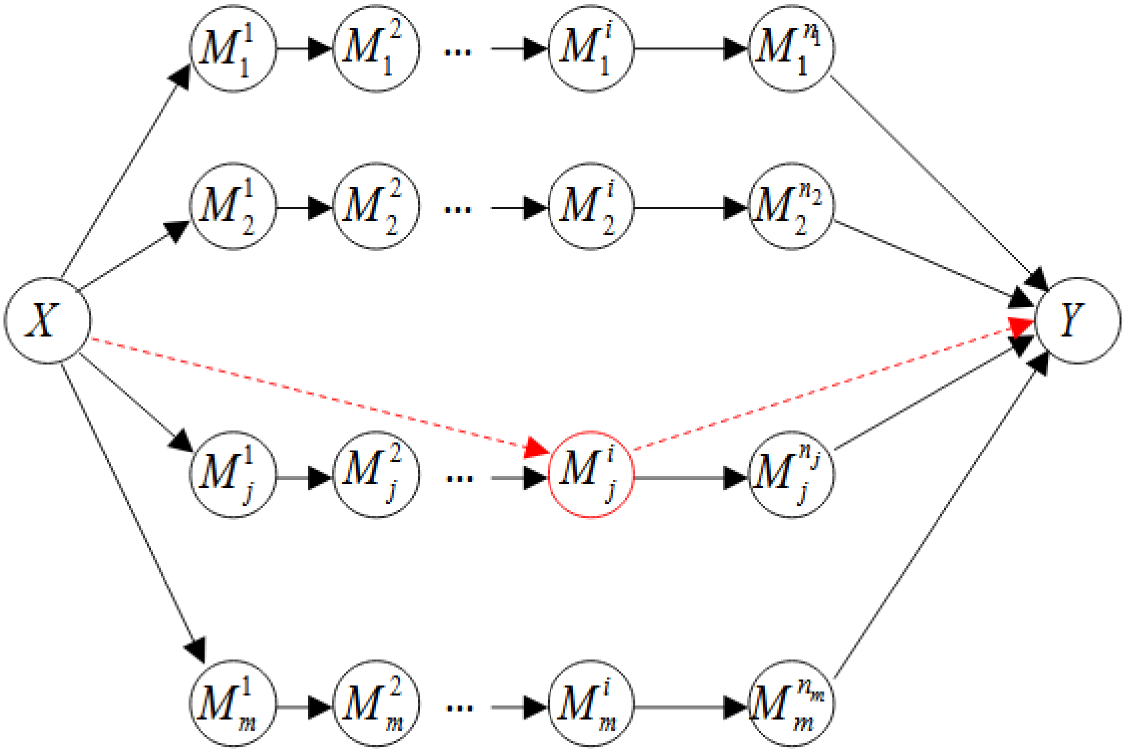

Figure 1: Example structure of a processing hierarchy in the brain.M
Here, X and Y encode the experimental manipulation and the ensuing behavioral response, respectively. Variables M_j^i are activity within brain regions that act as intermediary processing steps. In this oriented graphical model, arrows represent causal relationships. Although processing pathways operate both in series and in parallel, mass-univariate brain-behavior mediation analysis ignore this and treat each region/voxel independently of each other (red dotted arrows).

In simple bottom-up hierarchical architectures such as this one, lower levels would correspond to e.g., occipital low-level visual processes, whereas higher levels would map to, e.g., prefrontal decision making processes. Clearly, Figure 1 is already an oversimplification because it ignores reciprocal connections, branching processes and/or context-dependent gating mechanisms (Friston, 2011; He and Evans, 2010; Rubinov and Sporns, 2010). But when we perform mass-univariate brain-behavior mediation analysis, we reduce the complexity even further by considering each voxel or region of interest in isolation, effectively ignoring any hierarchical structure of this sort.

First, given the likely parallel nature of processing pathways, one would not expect that any isolated voxel or region of interest may ever *fully* mediate the impact of X onto Y. The implicit assumption of mass-univariate brain-behavior mediation analysis is that, in each voxel, the direct path c' effectively captures, in a non specific manner, mediated effects that go through other (parallel) pathways. This, however, places a very heavy load on the statistical sensitivity of mediation tests, which need to be able to detect potentially small indirect effect sizes, even when correcting for multiple comparisons (e.g., across voxels).

Second, nothing prevents different processing pathways to have strong but opposing impacts on the behavioral response. An example here would be opponent brain systems that yield strong but ambivalent (e.g., appetitive-aversive) cognitive states, whose idiosyncratic balance may explain one's specific ability to suppress e.g., impulsive behavioral responses (Seymour et al., 2005; Zhang et al., 2017). In particular, if the impact of different pathways balance out, then the total effect of X onto Y may become undetectable ($c \approx 0$). It follows that brain-behavior mediation analyses may be required for faithfully identifying the determinants of behavior. Alternatively, the relative contribution of different pathways may vary across individuals, which may drive inter-individual behavioral differences. We will see an example of this in the Results section below.

Statistical tests of mediation

In what follows, we recall the most established approaches to null-hypothesis testing of mediated effects. We start with the premise that if M mediates the effect of X onto Y, then the corresponding indirect effect has to be different from 0 ($ab \neq 0$). In what follows, we summarize two kinds of statistical testing approaches (namely: the “indirect” and the “conjunctive” approaches) that differ in terms of how they frame the corresponding null hypothesis.

The *indirect* approach follows from noting that the null hypothesis of mediation analysis can be framed as follows: $H_0^{(indirect)}: ab = 0$.

Under the simple brain-behavior mediation model in Equations 1-2, the indirect effect equates the difference between total and direct effects, i.e. $ab = c - c'$. This is why early approaches to mediation testing were assessing the statistical significance of the difference $c - c'$ (Baron and Kenny, 1986). However, theoretical work demonstrated that this equivalence may not always hold (Pearl, 2012), which would render the ensuing test invalid. This applies to typical fMRI experiments, because of the effect of confounding variables on path coefficient estimates.

Another, more valid, approach is to compare estimates of the indirect effect to their distribution under the null. This is the principle of *Sobel’s test* (Sobel, 1982). Recall that all parameters are identifiable from Equation 2, given X, Y and M. In particular, the ordinary least-squares (OLS) estimates $\hat{a}$ and $\hat{b}$ of unknown path coefficients a and b are given by (all X, Y and M variables are z-scored, see Appendix A for a mathematical derivation):

$$\begin{cases} \hat{a} = X^T M / n \\ \hat{b} = \dfrac{1}{\hat{\sigma}^2_{M|X}} \hat{\varepsilon}_M^{0\,T} Y / n \end{cases} \tag{4}$$

where the neural noise estimate $\hat{\varepsilon}_M^0 = M - X\hat{a}$ is the component of M that cannot be explained by X, and $\hat{\sigma}^2_{M|X} = 1 - \hat{a}^2$ is its sample variance. From Equation 4, one can see that $\hat{b}$ is simply the sample correlation between behavioral data Y and the neural noise estimate $\hat{\varepsilon}_M^0$. In other words, $\hat{b} \neq 0$ when M has an effect on Y *above and beyond the effect of* X. In addition, the variance of these OLS estimates are given by:

$$\begin{cases} \hat{\sigma}_a^2 = \dfrac{\hat{\sigma}_{M|X}^2}{n} \\ \hat{\sigma}_b^2 = \dfrac{\hat{\sigma}_{Y|M,X}^2}{(n-1)\hat{\sigma}_{M|X}^2} \end{cases} \tag{5}$$

where $\hat{\sigma}_{Y|M,X}^2 = 1 - \hat{b}^2 - \left(X^T Y/n - \hat{a}\hat{b}\right)^2$ is the sample variance of behavioral residuals' estimates $\hat{\varepsilon}_Y = Y - M\hat{b} + X\hat{c}'$.

Under the assumption that model residuals ε_Y and ε_M^0 are i.i.d. normal variables, then both $\hat{a}$ and $\hat{b}$ follow normal distributions: $\hat{a} \sim N\left(a, \hat{\sigma}_a^2\right)$ and $\hat{b} \sim N\left(b, \hat{\sigma}_b^2\right)$. It can then be shown (see Appendix B) that the product $\hat{a}\hat{b}$ approximately follows a normal distribution, i.e.: $\hat{a}\hat{b} \sim N\left(ab, \hat{\sigma}_a^2\hat{b}^2 + \hat{\sigma}_b^2\hat{a}^2\right)$. This implies that, under the null, the following pseudo z-statistics:

$$z_{ab}^{(Sobel)} = \frac{\hat{a}\,\hat{b}}{\sqrt{\hat{\sigma}_a^2\hat{b}^2 + \hat{\sigma}_b^2\hat{a}^2}} \tag{6}$$

approximately follows a Student probability density function. This then serves to derive the p-value of Sobel's unsigned (two-tailed) significance test $p_0^{(Sobel)} = 1 - 2\Phi\left(\left|z_{ab}^{(Sobel)}\right|\right)$, where Φ is Student's cumulative density function with appropriate degrees of freedom. Later improvements over Sobel's test (Hayes and Scharkow, 2013) derived from theoretical statistical works on the distribution of the product of two normal random variables, which essentially include an additional $\pm\hat{\sigma}_a^2\hat{\sigma}_b^2$ term to the denominator of Sobel's pseudo-zscore (Aroian, 1947; Goodman, 1960).

$$z_{ab}^{(Aroian)} = \frac{\hat{a}\hat{b}}{\sqrt{\hat{\sigma}_a^2\hat{b}^2 + \hat{\sigma}_b^2\hat{a}^2 + \hat{\sigma}_a^2\hat{\sigma}_b^2}}$$
$$z_{ab}^{(Goodman)} = \frac{\hat{a}\hat{b}}{\sqrt{\hat{\sigma}_a^2\hat{b}^2 + \hat{\sigma}_b^2\hat{a}^2 - \hat{\sigma}_a^2\hat{\sigma}_b^2}} \qquad (7)$$

We refer to these extensions as Aroian's and Goodman's tests, respectively. Alternatively, non-parametric approaches have been proposed to derive the distribution of indirect effect size estimates under the null (MacKinnon et al., 2002, 2004). Here, we will use the same bias-corrected bootstrap approach as the one proposed in the M3 toolbox (Wager, 2008).

The *conjunctive* approach follows from noticing that the null hypothesis of mediation analysis is composite (Moran, 1970), i.e.: $H_0^{(conjunction)}: a=0\,OR\,b=0$. Of course, both null hypotheses are exactly equivalent, but the composite null highlights the fact there is no mediated effect as long as one path coefficient is null (which breaks the causal cascade). In turn, one may test for the conjunction of both effects, i.e. test for the statistical significance of both a and b path coefficients. In practice, *conjunctive* testing relies on the "maximum p-value" approach (here, two-tailed test):

$$\begin{aligned} p_0^{(conj)} &= \max\left(2\Phi\left(-|t_a|\right), 2\Phi\left(-|t_b|\right)\right) \\ &= 1-2\Phi\left(\min\left(|t_a|,|t_b|\right)\right) \end{aligned} \qquad (8)$$

where $t_a = \hat{a}/\hat{\sigma}_a$ and $t_b = \hat{b}/\hat{\sigma}_b$ are Student's test statistics of a and b path coefficients, respectively. Formally speaking, $p_0^{(conj)}$ provides an *upper bound* on the joint probability that, under the null, two independent Student's test statistics take more extreme values than t_a and t_b (Friston et al., 2005a; Nichols et al., 2005). This is important, because conjunctive testing cannot be invalid but may have low sensitivity. However, it is trivial to show that Sobel's pseudo z-score is always smaller than the conjunctive test statistics, i.e.: $\left|z_{ab}^{(Sobel)}\right| \leq z_{ab}^{(conj)}$, where $z_{ab}^{(conj)} = \min\left[\left|t_a\right|, \left|t_b\right|\right]$ is the conjunctive test statistics (see Appendix B). This means that one would expect conjunctive testing to be systematically more efficient than Sobel's approach. At this point, we note that the sensitivity profile of indirect and conjunctive approaches actually depends upon neural noise strength and model misspecifications (see next sections). We will address this and related issues in the Results section, using extensive numerical Monte-Carlo simulations.

We refer the reader interested in extending these statistical approaches to experimental designs including multiple conditions (cf., e.g., factorial designs) and/or multiples subjects (cf. group-level random effects analysis) to Appendix C and D, respectively.

The non-trivial impact of neural noise

Although indirect and conjunctive null hypotheses are formally equivalent to each other, the latter is helpful to disclose the subtle tension behind mediation testing. In brief, two conditions must be satisfied for detecting a mediated effect: (i) strong evidence for $a \neq 0$ and (ii) strong evidence for $b \neq 0$. The former means that X partly explains the trial-by-trial variability of M. And the latter means that M partly explains the variability of Y that is unexplained by X. The critical point here is to realize that these two conditions are in conflict with each other. This is because they have opposing demands on neural noise ε_M^0. Note that the conjunctive test statistics $z_{ab}^{(conj)}$ is given by:

$$z_{ab}^{(conj)} = \min\left[\sqrt{n}\frac{|\hat{a}|}{\hat{\sigma}_{M|X}}, \sqrt{n-1}\frac{|\hat{b}|}{\hat{\sigma}_{Y|M,X}}\hat{\sigma}_{M|X}\right] \tag{9}$$

where we simply have inserted Equation 5 into the definition of the conjunctive test statistics. One can see that the standard deviation $\hat{\sigma}_{M|X}$ of the neural noise estimate $\hat{\varepsilon}_M^0$ will have opposing effects on the conjunctive test statistics. In brief, if $\hat{\sigma}_{M|X} \to 0$, then $z_{ab}^{(conj)} = |t_b|$, which tends towards 0 when $\hat{\sigma}_{M|X} \to 0$. Recall that, by definition, ε_M^0 is the component of M that cannot be explained by X (cf. Equation 2). Thus, in the absence of neural noise, the evidence for $a \neq 0$ is maximal, but M cannot explain any variability in Y that is unexplained by X, i.e. the evidence for $b \neq 0$ is minimal. Reciprocally, if $\hat{\sigma}_{M|X} \to \infty$, then $z_{ab}^{(conj)} = |t_a|$, which tends also towards 0 when $\hat{\sigma}_{M|X} \to \infty$. In other words, if neural noise strength is very high, then evidence for $a \neq 0$ is weak. Only for *intermediary levels of neural noise* can evidence for both $a \neq 0$ and $b \neq 0$ reach statistical significance. We note that this observation generalizes to any mediation test, irrespective of the mathematical form of the brain-mediation model. We refer the interested reader to Appendix E.

We will quantify the impact of neural noise on the statistical efficiency of candidate mediation testing approaches in the Results section below. But this property of mediation analysis has an important implication, which we now highlight.

Recall the structure of the processing hierarchy in Figure 2. Within a given processing pathway, each hierarchical level responds to its (lower-level) parents, eventually changing the information content in an incremental manner, e.g.:

$$\begin{cases} M_1 = Xa_0 + \varepsilon_M^0 \\ M_2 = M_1 a_1 + \varepsilon_M^1 \\ \quad \cdots \\ M_{i+1} = M_i a_i + \varepsilon_M^i \\ \quad \cdots \\ Y = M_N b + Xc' + \varepsilon_Y \end{cases} \tag{10}$$

where $\{M_1, M_2, \ldots, M_i, \ldots, M_N\}$ are local neural responses (indexed by their level along the hierarchy), and local neural noise increments ε_M^i effectively capture, in an agnostic manner, the unique contribution of each hierarchical level. Here, one would expect that local neural responses gradually diverge from the initial explanatory variable X. This is simply because the correlation between X and the local neural response M_i degrades as the accumulated neural noise increments $\sum_{j=0}^{i} \varepsilon_M^j$ increases. In turn, one would expect that mass-univariate mediation analysis can only detect those neural information processing steps that are positioned at an intermediary hierarchical level, i.e. sufficiently far away from either end of the hierarchy. We will exemplify this in the Results section below.

Dealing with hemodynamic confounds

Clearly, the brain-behavior mediation model in Equation 2 cannot directly be applied to fMRI time series. The reason is twofold. First, behavioral and neural variables are not sampled in the same manner. In brief, the former is collected at each "trial" of the behavioral task, while the latter is typically sampled at a sub-trial temporal resolution. Second, fMRI BOLD dynamics effectively result from the convolution of neural activity with the hemodynamic response function or HRF (Logothetis et al., 2001; Martin et al., 2006). This implies that the event-related BOLD response is delayed in time, when compared to trial onsets. In addition, if the inter-trial interval is smaller than the HRF duration (which is typically the case), BOLD signals measured during a trial may derive from the additive contributions of multiple neural responses (to the current and preceding trials). For the purpose of brain-behavior mediation analysis, there are essentially two ways of dealing with such hemodynamic confounds.

On the one hand, one may deconvolve BOLD signals from the HRF, as follows. Let τ_k (resp., Δ_k) be the onset time (resp., duration) of the k th trial in the experimental design. One first construct "trial" regressors that span the duration of the fMRI session (at the sampling resolution of fMRI ; typically: TR=1-2secs), which are zero everywhere except during the time interval defined as $\left[\tau_k, \tau_k + \Delta_k\right]$. Each of these is then convolved with the canonical HRF and its temporal derivatives, to account for potential mismatches in hemodynamic delays (Liao et al., 2002). One then augment the resulting GLM with fMRI confounds (e.g., motion regressors and slow drifts), and fit it to fMRI time series. Fitted regressor weights at each voxel thus provide an estimate $\hat{M}$ of the local neural response to each trial, which is deconvolved from the HRF and corrected for typical fMRI confounds, and can then enter a mediation analysis. We call this the *deconvolution* approach.

On the other hand, one may reframe the brain-behavior mediation model in the HRF-convolved space. One first resample the explanatory and dependent variables at the fMRI temporal resolution by reweighing each "trial'" regressor above with its corresponding X and Y entries and them summing over trials. One then convolves the resulting regressors with the canonical HRF (and its temporal derivatives) and augment the resulting GLM with fMRI confounds prior to entering a mediation analysis. We call this the *convolution* approach.

Both approaches can, in principle, deal with hemodynamic and other fMRI confounds, but they differ in terms of their respective bias-variance tradeoff. The *convolution* approach effectively yields reliable neural response estimates, under the implicit assumption that the HRF is identical across trials. In contrast, the *deconvolution* approach allows for trial-by-trial variations in HRF, at the cost of compromising the reliability of neural response estimates.

In the Results section below, we evaluate the robustness of these two strategies w.r.t. deviations to canonical HRF models.

A note on causality

Let us now highlight a possible interpretational issue of mediation analysis. Note that Equation 2 implicitly assumes a cascade of causal influences (MacKinnon et al., 2002), which may be best summarized in terms of the directed acyclic graph depicted on Figure 2 below (left panel).

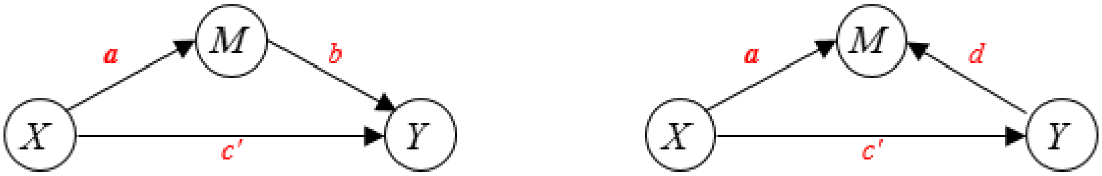

Figure 2: The two causal interpretations of mediated effects.
Left panel: “native” causal interpretation of brain-behavior mediation analysis (cf. Equation 2). Right panel: “swapped” causal interpretation of brain-behavior mediation analysis (cf. Equation 11). Corresponding path coefficients are shown in red.

One would then be tempted to interpret a statistically significant mediated effect in causal terms, as in: perturbing the independent variable X should result in changes in the mediator variable M that would eventually cascade down to the dependent variable Y. In the context of brain-behavior mediation, this causal interpretation aligns with the intuitive notion that behavioral responses to stimuli necessarily has to emerge from an intermediate neural information processing step. This causal reasoning, however, does not hold regarding the relationship between M and Y, which are both *observed* data. In Equation 2, the strength of this relationship is controlled by the path coefficient b. Importantly, statistical inference on the path coefficient b is but a quantitative assessment of the conditional mutual information $I\left(M,Y|X\right)$, which is invariant under a reversal of the directionality of the relationship between M and Y. In other terms, Equation 2 is formally equivalent to the following alternative model:

$$\begin{cases} M = Xa + \varepsilon_M^0 \\ M = Yd + Xa' + \varepsilon_M \end{cases} \tag{11}$$

where the second line simply derives from swapping (with impunity) the explanatory and response variables in the second line of Equation 2. Here, d and a' are "swapped" path coefficients that have a different causal interpretation (cf. Figure 1, right panel), and ε_M are model residuals that are not equivalent to the neural noise ε_M^0 of Equation 2. One can show (see Appendix E) that both native and swapped path coefficients estimates are analytically related as follows:

$$\hat{b} = \hat{d}\,\frac{\hat{\sigma}_{Y|X}^2}{\hat{\sigma}_{M|X}^2} \tag{12}$$

where $\hat{\sigma}^2_{Y|X} = 1 - \left(X^T Y / n \right)^2$ is the sample variance of Equation 1's residuals estimates $\hat{\varepsilon}^0_Y = \left(I - XX^T / n \right) Y$. It should be clear from Equations 11-12 that assessing the conditional mutual information $I\left(M, Y | X \right)$ can be equivalently addressed either by assessing the evidence for $b \neq 0$ (native form of the mediation model, cf. Equation 2), or by assessing the evidence for $d \neq 0$ (cf. Equation 11). In fact, the ensuing t-statistics t_b and t_d are exactly equal (see Appendix E), i.e. brain-behavior mediation test statistics are invariant under a permutation of M and Y variables.

This has two important consequences.

First, one may rely on Equations 11-12 to improve the computational efficiency of brain-behavior mediation analysis by several orders of magnitude. Recall that in the context of whole-brain fMRI, working with regression models where fMRI signals only enter as dependant variables is computationally very advantageous. This is because many algebraic operations that are required for parameter estimation (e.g., here, matrix multiplications and inversions, etc) can be computed once and for all. In brief, the computational gain of performing brain-behavior mediation analysis using Equations 9-10, when compared to Equation 2, is of the order of $n^2_{scan} \times n_{voxel}$, where n_{scan} and n_{voxel} are the number of fMRI time samples and voxels, respectively. This may speed up whole-brain mediation analysis by several orders of magnitude.

Second, a statistically significant mediated effect is compatible with two causal interpretations. In particular, under the "swapped" model of Equation 11, variations in behavior Y may cause changes in the neural response M (cf. Figure 1, right panel).

This alternative causal interpretation ($Y \to M$) is not as nonsensical as it may first sound. For example, somatosensory cortices will respond to variations in motor actions, eventually enabling proprioceptive sensations. More generally, a given brain system may be collecting and/or processing information regarding overt behavior (which may have been produced elsewhere in the brain) for the purpose of, e.g., learning, memory, metacognition, etc... In any case, this interpretational issue is important, because the implicit intention behind brain-behavior mediation analysis is clearly to provide statistical evidence for the "native" causal scenario ($X \to M \to Y$). We will comment on this and related issues in the Discussion section of this manuscript.

One may think that affording evidence for the "native" causal claim of brain-behavior mediation analysis may require non observational studies, e.g., causal perturbations of neural activity (lesion studies, transcranial magnetic stimulation, etc). Nevertheless, we argue that one may perform complementary data analyses that may partially address the interpretational issue above. For example, having assessed the significance of a mediated effect, one may exploit locally multivariate information to provide statistical evidence for or against candidate causal claims. In fact, when considering the set of mediator variables within a significant cluster together, "native" ($M \to Y$) and “swapped” ($Y \to M$) causal interpretations induce a many-to-one and a one-to-many M-Y mapping, respectively (see Figure 3 below). Because "native" and "swapped" causal scenarios differ in terms of whether Y is viewed as an input or as an output of local neural information processing, we refer to the ensuing test statistics as an *I/O test statistics*.

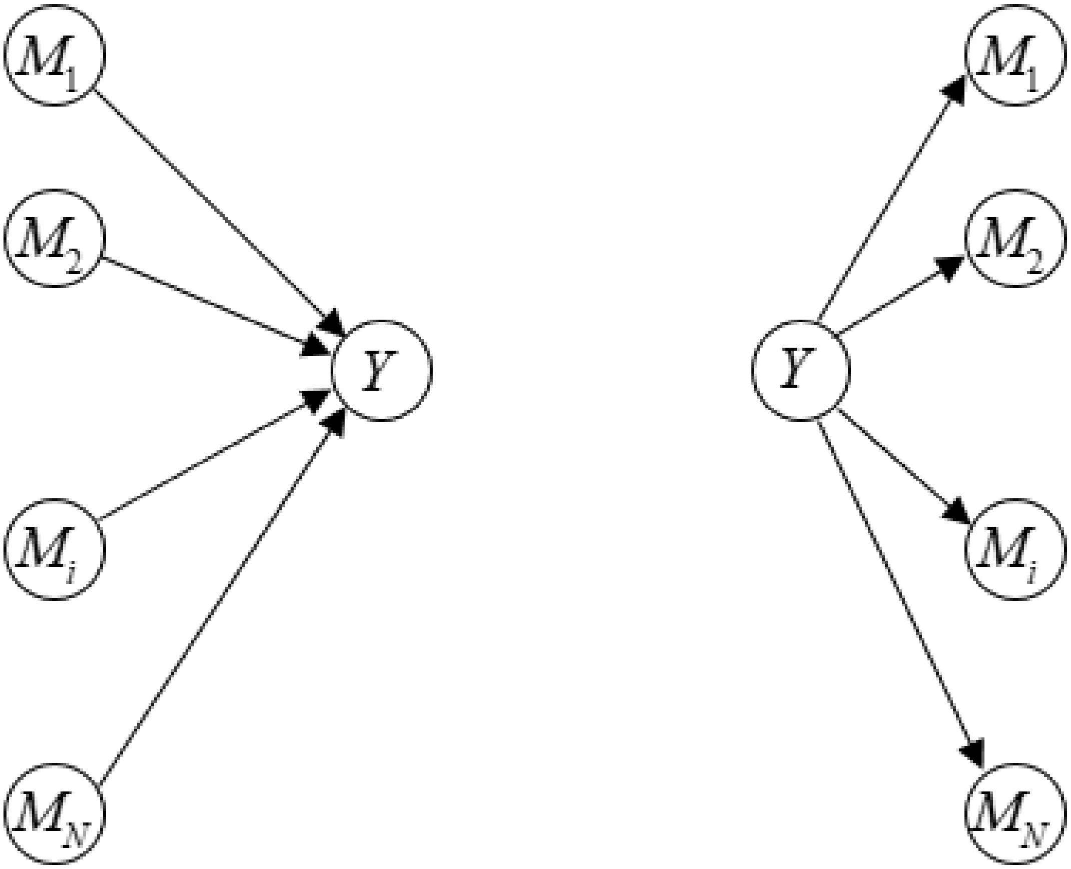

Figure 3: Candidate multivariate input/output M-Y mappings.
Left panel: “native” causal interpretation (many-to-one input/output mapping, Y as an output). Right panel: “swapped” causal interpretation (one-to-many input/output mapping, Y as an input).

Let M_i be the trial-by-trial variations of a voxel belonging to a given mediator cluster, where $i \in [1, N]$ and N is the number of voxels in the cluster. We define our I/O test statistics $\bar{\lambda}$ as follows:

$$\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i \qquad (13)$$

where λ_i is the loss of conditional mutual information between M_i and Y when accounting for other neighboring voxels $M_{j\neq i}$:

$$\lambda_i = I\left(M_i, Y \middle| X\right) - I\left(M_i, Y \middle| X, M_{j\neq i}\right) \tag{14}$$

In Equation 14, $I\left(M_i, Y \middle| X, M_{j\neq i}\right)$ and $I\left(M_i, Y \middle| X\right)$ are the conditional mutual information between M_i and Y, given X and the activity in all other voxels $j\neq i$ or not, respectively. Note that λ_i is sometimes coined the *interaction information* (McGill, 1954). In brief, $\bar{\lambda}$ measures the average improvement or worsening of the mutual information between candidate mediator voxels and behavioral responses, when accounting for variations in neighboring brain activity.

For arbitrary gaussian variables X and Y, the mutual information $I(X,Y)$ can be written as $I(X,Y) = -1/2\log\left(1-\rho_{X,Y}{}^2\right)$, where $\rho_{X,Y}$ is the correlation between X and Y (Marrelec et al., 2005). In turn, Equation 12 can be rewritten as follows:

$$\bar{\lambda} = -\frac{1}{2N}\sum_{i=1}^{N}\log\left(\frac{1-\hat{b}_i^2}{1-\tilde{b}_i^2}\right) \tag{15}$$

where $\tilde{b}_i$ is the conditional correlation between M_i and Y, given X and $M_{j\neq i}$:

$$\tilde{b}_i = \frac{\tilde{Y}_i^T\tilde{M}_i}{\sqrt{\tilde{Y}_i^T\tilde{Y}_i}\sqrt{\tilde{M}_i{}^T\tilde{M}_i}} \quad \text{with} \quad \begin{cases} Z_i = \left[M_{j\neq i}, X\right] \\ P_i = I - Z_i\left(Z_i{}^T Z_i\right)^{-1} Z_i{}^T \\ \tilde{Y}_i = P_i Y \\ \tilde{M}_i = P_i M_i \end{cases} \tag{16}$$

It turns out that the sign of $\bar{\lambda}$ provides evidence in favor or against the native causal interpretation of the brain-behavior mediation model. More precisely: if $M \to Y$, then $E\left[\bar{\lambda}\right] \leq 0$, whereas if $Y \to M$, then $E\left[\bar{\lambda}\right] \geq 0$. This is because if Y is an output of local brain activity ($M \rightarrow Y$), then any given univariate statistical relationship between M_i and Y (path coefficient $\hat{b}_i$) is obscured by the (partially independent) contributions of all other mediator variables $M_{j\neq i}$. Therefore, when removing all the variability that can be explained with $M_{j\neq i}$, one reveals the unique contribution of M_i (i.e. $\hat{b}_i < \tilde{b}_i$). In contrast, if Y is an input to local brain activity ($Y \rightarrow M$), then the variability shared by all mediator variables results from the influence of Y. Therefore, when removing all the variability that can be explained with $M_{j\neq i}$, one degrades the statistical relationship between M_i and Y (i.e. $\hat{b}_i > \tilde{b}_i$). We will evaluate the utility and robustness of our I/O test statistics in the Results section below.

Results

In what follows, we will be comparing five testing approaches: Sobel's test, Aorian's test, Goodman's test, the M3 bootstrap test, and the conjunctive approach, in terms of their statistical sensitivity and specificity. Using numerical simulations, we will assess the impact of neural noise and deviations to HRF assumptions. Taken together, these *in-silico* experiments will serve to address questions Q1 to Q3. Using further numerical simulations, we will demonstrate the utility of our I/O test statistics for addressing the main interpretational issue of brain-behavior mediation analysis (Q4). Finally, we will report the results of a brain-behavior mediation analysis in the context of an fMRI experiment on decision making under risk.

Comparing the statistical specificity and sensitivity of testing approaches

First, we ask whether candidate testing approaches yield valid inferences, i.e. whether they allow for a faithful control of false positive rate. To address this question, we simulated data under three different variants of the null hypothesis. More precisely, we simulated 40,000 datasets with Equation 2, using three different settings of the path coefficients, i.e.: (i) $a=0$ and $b=1/2$, (ii) $a=1/2$ and $b=0$, or (iii) $a=b=0$. In all simulations, we simulated $n=50$ trials, set the direct effect size to $c'=1/2$ and used unitary variance for all independent variables in Equation 2 (i.e. X, ε_M^0 and ε_Y). Across these 40,000 simulations, we then measured the (false positive) detection rate of each candidate testing approach, as one varies the significance threshold α. Note that all (indirect or conjunctive) parametric tests were performed with Student's probability distribution functions with $n-2$ degrees of freedom. Finally, we kept the default number of 1000 resamplings in the bias-corrected M3 bootstrap test.

Second, we asked how sensitive are candidate testing approaches under moderate mediated effect sizes. Here, we simulated 40,000 datasets with Equation 2, using $a=b=1/2$, and measured the (true positive) detection rate of each candidate testing approach, as one varies the significance threshold α.

The results of these analyses are summarized on Figure 4 below.

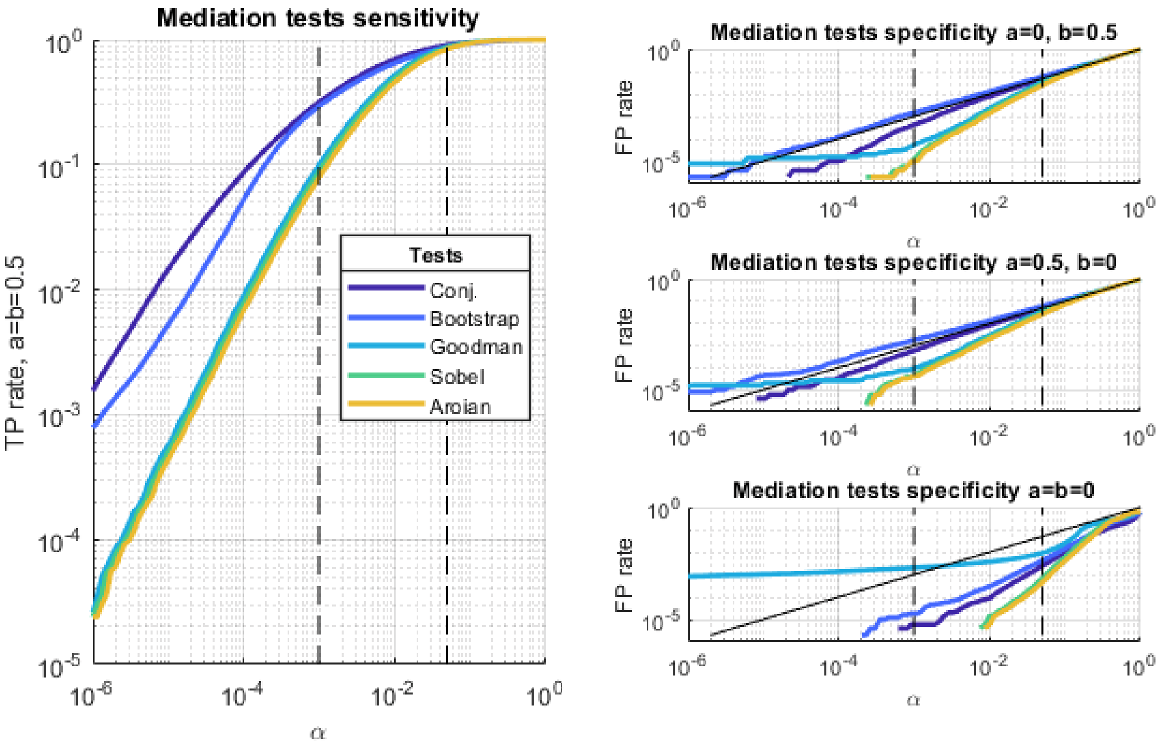

Figure 4: Statistical specificity and sensitivity of variants of mediation significance testing approaches.

Left panel: The sensitivity of mediation tests (y-axis) is plotted against the significance threshold α (x-axis), for each candidate testing approach (dark blue: conjunctive testing, blue: M3 bootstrap indirect approach, light blue: Goodman's indirect approach, green: Sobel's indirect approach, yellow: Aorian's indirect approach). Upper right panel: The specificity of mediation tests (y-axis) is plotted against the significance threshold, for H_0: a=0 and b=1/2 (same format as left panel). Middle right panel: Same format as upper right panel for H_0: a=1/2 and b=0. Lower right panel: Same format as upper right panel for H_0: a=b=0.

As expected, the conjunctive test is more sensitive than both Sobel and Aorian tests. Slightly more surprising maybe is the fact that the conjunctive test turns out to also be more sensitive than Goodmans test and the M3 bootstrap test, though the latter reach similar sensitivity levels for significance thresholds higher than 0.001. We will refine our evaluation of statistical sensitivity when assessing the impact of neural noise below.

In addition, all approaches except Goodman and the M3 bootstrap tests are valid, i.e. they yield a false positive rate that is equal or smaller than the significance threshold α. Goodman's test always yield invalid inference if the significance threshold is small enough, whereas the M3 bootstrap test only yields invalid inference when $b=0$. Note that the conjunctive approach is the least conservative of all tests, and this difference grows when the significance threshold decreases.

Assessing the impact of neural noise

Recall that the magnitude of neural noise is expected to play a critical role for the statistical sensitivity of mediation analysis. To demonstrate this effect, we simulated 10,000 datasets using the same parameter settings as above, except for neural noise magnitude, which we varied from $Var\left[\varepsilon_M^0\right]=10^{-2}$ to $Var\left[\varepsilon_M^0\right]=10^{2}$. For each neural noise magnitude, we then measured the (true positive) detection rate of each candidate testing approach, when setting the significance threshold to $\alpha=0.05$. The ensuing sensitivity profiles are summarized on Figure 5 below (left panel).

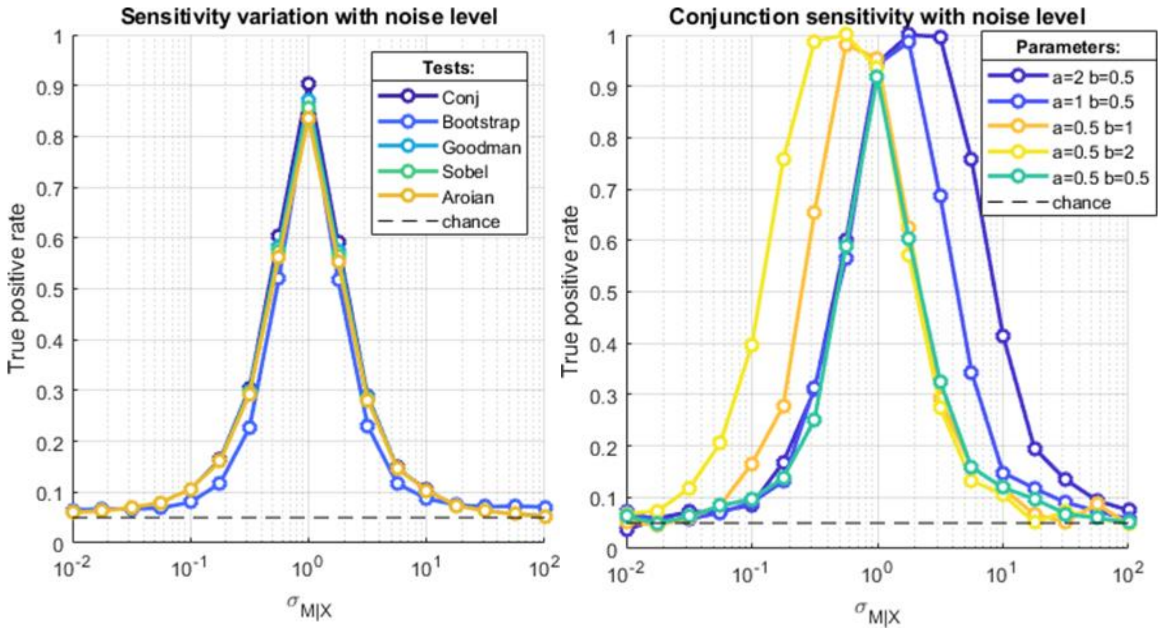

Figure 5: The impact of neural noise on statistical power.
Left panel: The sensitivity of mediation tests (y-axis) is plotted against the variance of neural noise (x-axis), for each candidate testing approach (same format as Figure 4), when a=b=1/2. Chance level is indicted using a black dotted line. Right panel: The sensitivity of conjunctive

mediation tests (y-axis) is plotted against the variance of neural noise (x-axis), when varying path coefficients (dark blue: a=2 and b=1/2, blue: a=1 and b=1/2, cyan: a=b=1/2, orange: a=1/2 and b=1, yellow: a=1/2 and b=2).

All testing approaches have a similar sensitivity profile, which follows a bell-shaped function of neural noise magnitude, with an apex around $Var\left[\varepsilon_M^0\right]=1$. This corresponds to a situation in which about 20% of the trial-by-trial variance in M is explained by X. As the amount of explained variance in M departs from this nominal level, the sensitivity of mediation analysis effectively tends towards chance level. Now, everything else being equal, increasing a or the variance of X eventually inflates sensitivity on the right tail of the sensitivity profile, while increasing b rather boosts its left tail. This moves the position of sensitivity apex towards smaller and stronger noise variance, respectively (see Figure 5, right panel).

Now, in the Methods section above, we reasoned that the expected sensitivity profile of mediation analysis should eventually favor the detection of neural information processing steps that are positioned away from either end of the processing hierarchy. In what follows, we compare candidate testing approaches w.r.t. their ability to detect levels in a simple feed-forward hierarchy. In brief, we simulated 1,000 datasets under Equation 10, using 100 intermediary network nodes. In all simulations, initial and final path coefficients were set to $a_0=b=1/2$ and all intermediary path coefficients were set to $a_i=1\,\forall i$. In addition, the variance of all independent variables were set to unity except for the local neural noise increments, whose standard deviation was set to 0.3. Following the principle of mass-univariate mediation analysis, a mediation test was then performed on each node in isolation (significance threshold: $\alpha=0.05$). For each network node, the ensuing (true positive) detection rate was then measured across the 1,000 simulations. The result of the ensuing detection profile is shown on Figure 6 below.

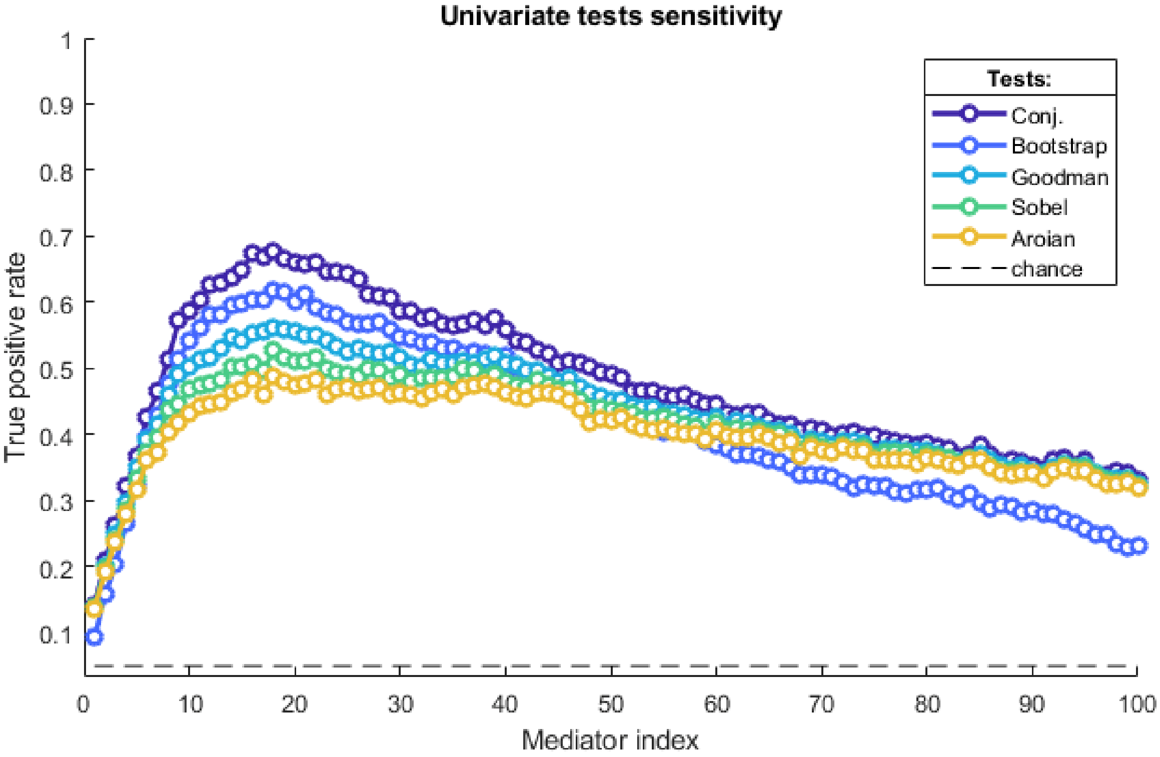

Figure 6: The heterogeneity of statistical sensitivity.
The sensitivity of mediation tests (y-axis) is plotted against the hierarchical level of candidate mediators along the serial processing pathway (x-axis), for each candidate testing approach (same format as Figure 4).

As expected, local neural noise increments accumulate along the hierarchy, effectively increasing the neural noise level estimate as the hierarchical level increases. In turn, the detection profile also follows a bell-shaped function of hierarchical level, such that lower and higher hierarchical levels are less easy to detect. Interestingly, one can also see that different testing approaches have different sensitivity profiles. In particular, one can see that the conjunctive approach exhibits a higher sensitivity than all other approaches, irrespective of the hierarchical level of interest. Note that the M3 bootstrap test is better than other indirect approaches for intermediate hierarchical levels, but eventually loses its competitive advantage for higher hierarchical levels.

Assessing the robustness to deviations from hemodynamic assumptions

Despite the inclusion of HRF derivatives in the mediation model, deviations to the canonical HRF can impair test sensitivity. In this section, we compare the robustness of *convolution* and *deconvolution* approaches to unanticipated delays in HRF. We thus simulated 100 datasets using the same parameter settings as above, except that we varied systematically the HRF delay, effectively inducing a shift with the canonical HRF ranging from -5 second to 5 seconds. Each dataset was then analyzed using all (*indirect* and *conjunctive*) testing approaches, under both *convolution* and *deconvolution* strategies (with the canonical HRF and its delay derivative). For each HRF delay shift, we then measured the (true positive) detection rate of each candidate mediation analysis strategy, when setting the significance threshold to $\alpha = 0.05$. The ensuing sensitivity profiles are summarized on Figure 7 below.

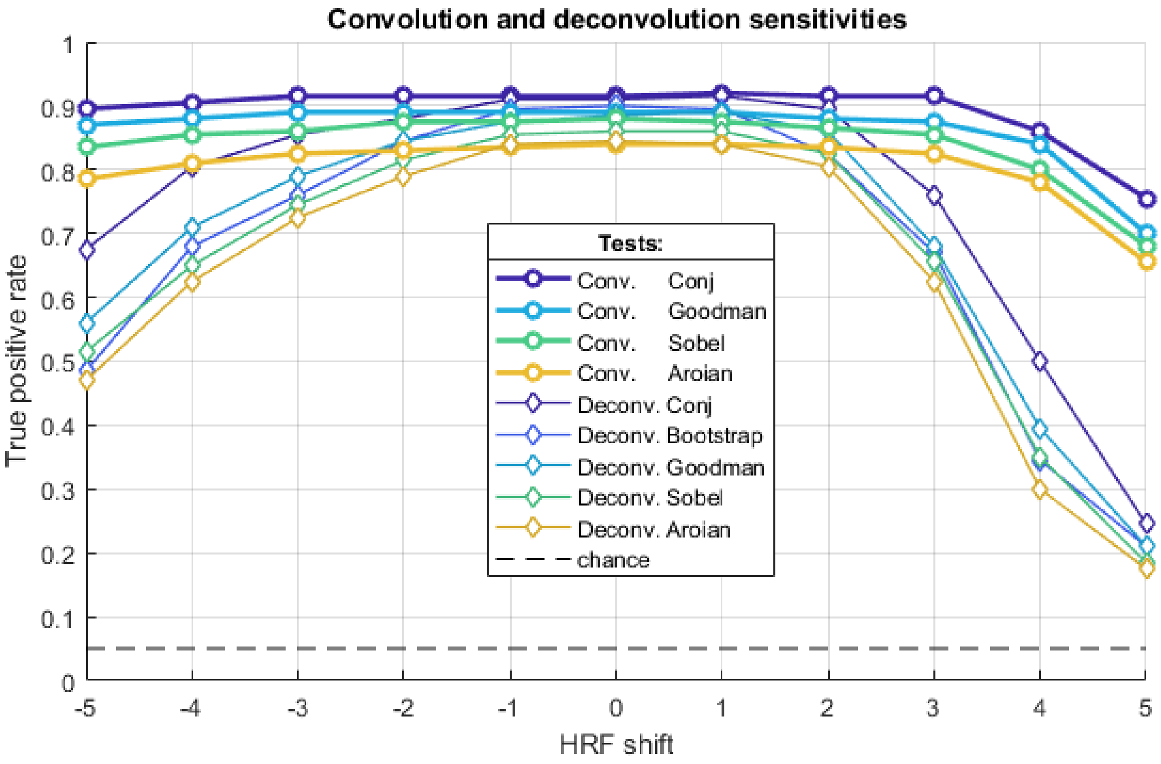

Figure 7: The impact of unmodelled hemodynamic delays.
The sensitivity of mediation tests (y-axis) is plotted against the HRF shift (x-axis), for each candidate testing approach (same format as Figure 4, thick lines: *convolution* approach, thin lines: *deconvolution* approaches).

All mediation analysis strategies exhibit a bell-shaped sensitivity profile, eventually peaking when there is no deviation to the canonical HRF (i.e. when the HRF delay shift is null). Also, when there is no deviation to the canonical HRF, *deconvolution* and *convolution* strategies yield similar test sensitivity. However, when the deviation to the canonical HRF increases, the loss of statistical sensitivity is much stronger for *deconvolution* than for *convolution* approaches. For example, with a (realistic) delay shift of 3 seconds, most *deconvolution* approaches lose about 10% to 15% sensitivity on average. In comparison, *convolution* approaches only lose about 2% sensitivity. In addition, the *conjunctive* approach always exhibit higher sensitivity than *indirect* approaches, irrespective of whether one chooses a *convolution* or *deconvolution* strategy.

We note that, with a significance threshold of $\alpha = 0.05$, deviations to the canonical HRF has no adverse effect on the validity of statistical tests, i.e. all mediation test approaches yield 5% or less false positive rates under the null.

Addressing the interpretational issue of brain-behavior mediation analysis with the I/O test statistics

Recall that a significant mediated effect may have two distinct causal interpretations: the behavioral variable may either be an input ($Y \to M$) or an output ($M \to Y$) of the brain region where the null has been rejected. To address this issue, we proposed a simple I/O test statistics, whose sign is expected to discriminate between these two scenarios.

Here, we evaluate the utility of the I/O test statistics $\bar{\lambda}$, in conditions similar to our fMRI data analysis below, using numerical Monte-Carlo simulations.

First, we simulated data under three scenarios:

- H_1 (native causal scenario $M \to Y$): the independent variable X is sampled under a normal distribution, each multivariate mediator unit M_i is set to a noisy affine transformation of X (with random weights), and the dependent variable Y is set as a noisy mixture of X and all mediator units (with random weights).
- H_2 (“swapped” causal scenario $Y \to M$): the independent variable X is sampled under a normal distribution, the dependent variable Y is set as a noisy affine transformation of X (with a random weight) and each multivariate mediator unit M_i is set to a noisy mixture of X and Y (with random weights).

- H_0 (null scenario): the independent variable X is sampled under a normal distribution, and all other variables are set to a noisy affine transformation of X (with random weights).

We simulated each scenario 1000 times, with 64 trials and 20 mediating units (all random variables and weights were sampled under a centered normal distribution with unit variance). Note that, in all three scenarios, M and Y variables are correlated with each other (under the null, this is because of the influence of X, which acts as a confounding variable). For each simulated dataset, we derive the I/O test statistics $\bar{\lambda}$. The resulting Monte-Carlo distributions are shown on Figure 8 below (left panel).

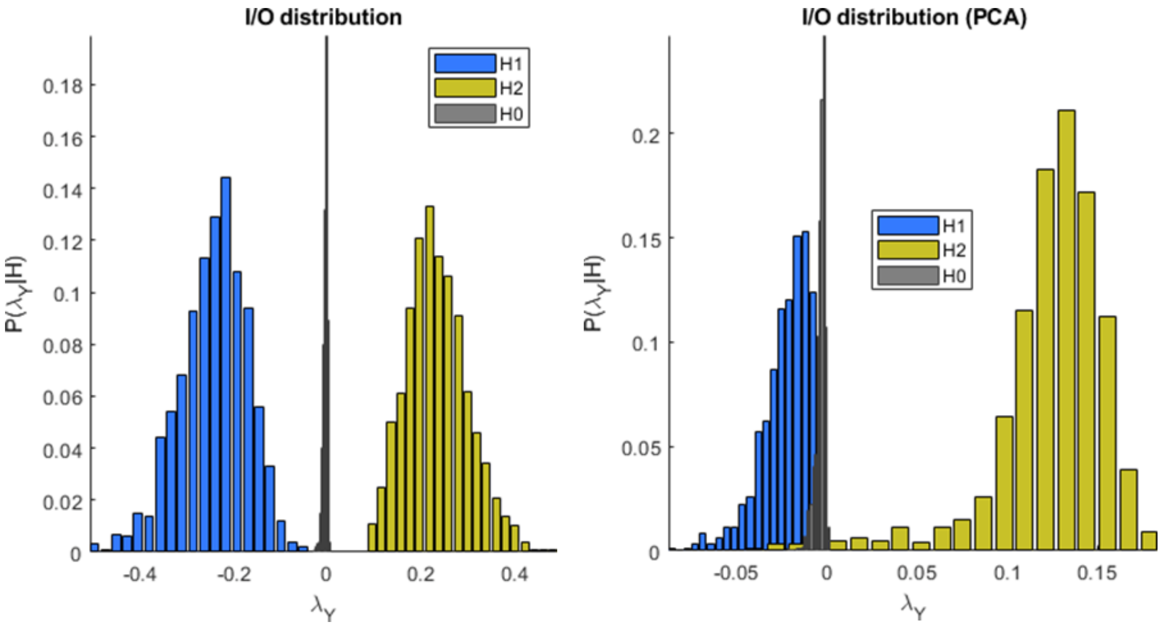

Figure 8 : Sensitivity and robustness of the I/O test statistics λ.
Left panel: The Monte-Carlo distribution of the I/O test statistics λ (y-axis) is plotted under alternative scenarios (H1: blue, H2: yellow, H0: grey). Right panel: Same format as left panel, but under data dimension reduction (20 first principal components of a PCA).

On can see that the three scenarios are very well discriminated. In particular, the distribution of the $\bar{\lambda}$ under the null is centered on zero, and lies in between its distribution under H$_1$ and under H$_2$. Moreover, and as expected, $E\left[\bar{\lambda}\middle|H_1\right]<0$ and $E\left[\bar{\lambda}\middle|H_2\right]>0$.

These simulations however, do not account for the limitations that arise in realistic settings. In particular, the number of neural units or voxels that compose the multivariate set of mediators may largely exceed the number of trials. Here, a pragmatic solution is to perform a PCA decomposition, and keep the *K* first principal components to summarize the within-region variability. We now ask whether the ensuing I/O test statistic is robust to this dimension reduction. In brief, we performed the same set of simulations as above, this time simulating 100 mediating units and deriving the test statistics from the ensuing *K*=20 first principal components. The resulting Monte-Carlo distributions are shown on the right panel of Figure 8.

One can see that the dimension reduction strongly reduces the range of variation of the I/O test statistics, when compared to the situation above, where all the relevant variation is available. Furthermore, the magnitude of $\bar{\lambda}$ under scenario H$_1$ and H$_2$ is asymmetrical. More precisely, one can see that $\left|E\left[\bar{\lambda}\middle|H_1\right]\right|<\left|E\left[\bar{\lambda}\middle|H_2\right]\right|$. In other terms, when relevant information is lacking, the average evidence in favor of H$_1$ is weaker than the average evidence in favor of H$_2$. Nevertheless, the sign of the I/O test statistics can still be interpreted as evidence for or against the native causal interpretation of the brain-behavior mediation model, i.e. $E\left[\bar{\lambda}\middle|H_1\right]<0$ and $E\left[\bar{\lambda}\middle|H_1\right]>0$.

fMRI study of decision making under risk

Here, we perform a brain-behavior mediation analysis of previously acquired fMRI data (Chen, 2014), which is openly available as part of the OpenFMRI project (Poldrack et al., 2013). In this study, 60 participants made a series of 64 accept/reject decisions on risky gambles. On each trial, a gamble was presented, entailing a 50/50 chance of gaining an amount G of money or losing an amount L (so-called "baseline" condition). Participants were told that, at the end of the experiment, four trials would be selected at random: for those trials in which they had accepted the corresponding gamble, the outcome would be decided with a coin toss, and for the other ones -if any-, the gamble would not be played. All 64 possible combinations of G/L pairs (10$<G<40$, 5$<L<20$) were presented across trials, which were separated by 7 seconds on average (min 6, max 10). MRI scanning was performed on a 3T Siemens Prisma scanner. High-resolution T1w structural images were acquired using a magnetization prepared rapid gradient echo (MPRAGE) pulse sequence with the following parameters: TR = 2530 ms, TE = 2.99 ms, FA = 7, FOV = 224 × 224 mm, resolution = 1 × 1 × 1 mm. Whole-brain fMRI data were acquired using echo-planar imaging with multi-band acceleration factor of 4 and parallel imaging factor (iPAT) of 2, TR = 1000 ms, TE = 30 ms, flip angle = 68 degrees, in-plane resolution of 2X2 mm 30 degrees of the anterior commissure-posterior commissure line to reduce the frontal signal dropout, with a slice thickness of 2 mm and a gap of 0.4 mm between slices to cover the entire brain. See https://openneuro.org/datasets/ds000053/versions/00001 for more details.

Data preprocessing included standard realignment and movement correction steps. Note that we excluded 2 participants, either due to missing information or because the misalignment between functional and anatomical scans could not be corrected.

We first regressed, for each participant, the observed choices against gains and losses (Equation 1). This yielded estimates of the total effects $\hat{c}_G$ and $\hat{c}_L$ of gains and losses, respectively. This also provided an estimate $\hat{\sigma}_{Y|X}$ of the behavioral residuals' standard deviation. The results of this analysis are shown on Figure 9 below.

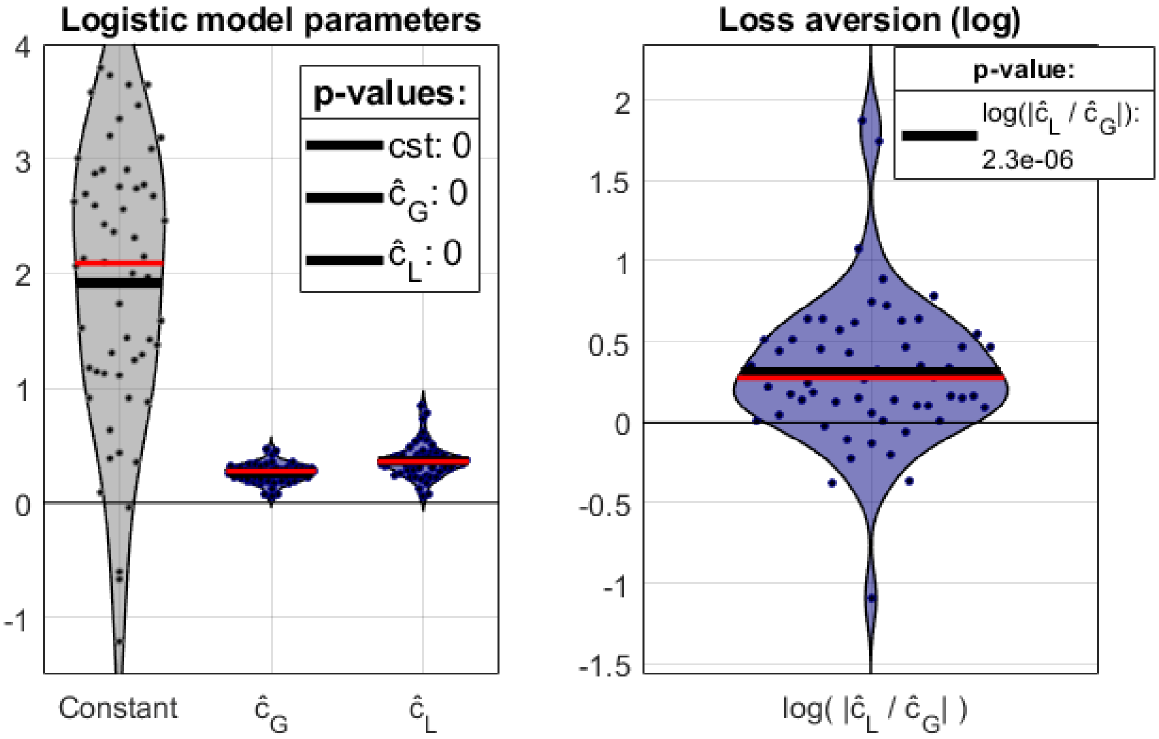

Figure 9: Summary of behavioral results.
Left panel: Between-subject empirical distribution of estimated within-subject parameters (left: constant term in the regression, middle: gain weight c_G, right: loss weight c_L). The black and red lines show the group-level mean and median, respectively. Right panel: Between-subject empirical distribution of the loss-version index (same format as left panel).

In brief, both gain and loss factors have a significant effect on decisions under risk (gain factor: $p<10^{-5}$, loss factor: $p<10^{-5}$). We note that, together, gain and loss factors explain on average 44.6% (std: 24.2%) of the trial-by-trial variance on participants’ decisions (average balanced accuracy: 84.84%, std: 9%).

For each participant, we also derived a loss-aversion index: $\omega = \log\left(\hat{c}_L/\hat{c}_G\right)$, which is positive when losses have a stronger weight on accept/reject decisions than gains. One can see that the average loss-aversion index is significant ($p<10^{-5}$), i.e. losses have more weight on participants' decisions than gains.

Then, we analyzed fMRI time series (at the within-subject level) using both *convolution* and *deconvolution* approaches.

The *convolution* strategy relied upon the following two GLMs:

- **Equation 10 (first line):** GLM1 included regressors for trial-by-trial gains and losses (temporally aligned with the gamble presentation and convolved with the canonical HRF and its delay derivative), and basic confounding factors (six movement regressors and their squared values, as well as a Fourier basis set for slow drift removal). Fitting GLM1 to each fMRI voxel time series yielded a map of estimates $\hat{a}_G$ and $\hat{a}_L$ that correspond to the local effect of gain and loss on neural activity at the time of gamble presentation, respectively. In addition, we extracted the standard deviation of GLM1's residuals, which form a map of the local neural noise's strength $\hat{\sigma}_{M|X}$.
- **Equation 11 (second line):** GLM2 is identical to GLM1, but also includes acceptance/rejection choices (convolved with the canonical HRF and its temporal derivatives). Fitting GLM2 to fMRI time series yielded regressor weight estimates that measure the correlation between local neural activity and behavior, above and beyond the effect of gain and losses ($\hat{d}$). The map of local path coefficients $\hat{b}$ was then obtained from $\hat{d}$, $\hat{\sigma}_{Y|X}$ and $\hat{\sigma}_{M|X}$ using Equation 12.

The *deconvolution* strategy was implemented as follows. First, we fitted GLM3, which included "trial" regressors (temporally aligned with the gamble presentation) as well as basic fMRI confounds. Regression weight estimates yielded local trial-by-trial neural responses $\hat{M}$. Maps of path coefficients estimates $\hat{a}$ and $\hat{b}$ were obtained using Equation 4, given local neural responses $\hat{M}$.

Random-effect group-level inference on the mediation of gain and loss factors was then performed by reporting group averages of path coefficients $\hat{a}$ and $\hat{b}$, after 8mm FWHM smoothing. We applied all indirect and conjunctive approaches except for the M3 bootstrap method (because of its limited statistical gain, when compared to its computational cost). For all approaches, we used unsigned (two-tailed) tests with standard random field theory (RFT) correction for whole-brain multiple comparisons correction.

In brief, no mediation testing approach based upon the *deconvolution* strategy reached statistical significance. This was the case even when using more lenient corrections for multiple comparisons (e.g., FDR). This was however not the case for mediation analyses based upon the *convolution* strategy. Here, *indirect* approaches yielded group-level significant clusters at low-set inducing thresholds (p=0.01 or p=0.05, uncorrected). In what follows, we discard these results as these thresholds are known to violate RFT assumptions (Flandin and Friston, 2019). Now, under the default set-inducing threshold (p=0.001, uncorrected), the conjunctive approach identified 6 clusters that significantly mediate the effect of gain: the right supramarginal gyrus or SMG (p=0.011, RFT-corrected), bilateral posterior dorsomedial PFC or BA8 (left: p=0.003, right: p=0.008, RFT-corrected), the right anterior ventrolateral PFC or BA45 (p=0.018, RFT-corrected) and bilateral posterior dorsolateral PFC or BA8/9 (left: p=0.007, right: p=0.009, RFT-corrected). In addition, there was a trend (p=0.06, RFT-corrected) for 1 cluster mediating the effect of loss, in the left anterior ventrolateral PFC. These clusters are shown on Figure 10 below.

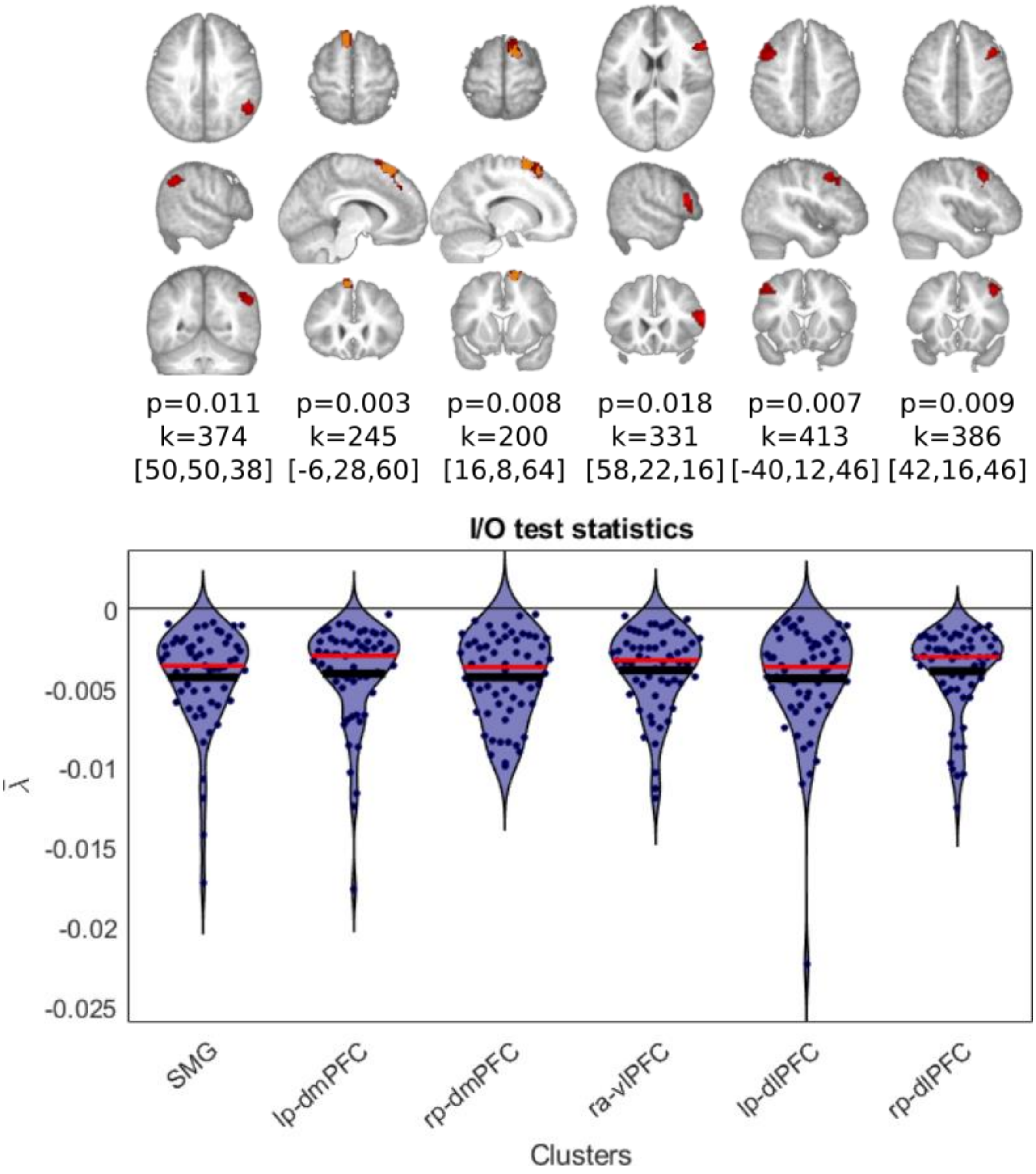

Figure 10: Significant mediators of gains and losses on decisions under risk.

Upper panel: the six significant clusters of brain-behavior mediation analysis (*conjunctive/convolution* approach) are shown on axial (up), sagittal (middle) and coronal views (bottom). All maps used a default set-inducing threshold of correction p=0.001 uncorrected (red areas) for the RFT correction, except the bilateral dmPFC's map where with p=0.0002 uncorrected (yellow areas) in order to separate the two hemispheres. Lower panel: The ensuing between-subject empirical distribution of the I/O test statistics λ (y-axis, group-level mean ±standard deviation) is shown for each significant clusters (x-axis).

We note that regions contralateral to significant unilateral mediators of the gain effect were all close to statistical significance: c.f. left SMG (p=0.094, RFT-corrected) and left anterior vlPFC or BA45 (p=0.177, RFT-corrected).

At the very least, these analyses demonstrate the superior statistical efficiency of *conjunctive/convolution* approaches. In brief, no other candidate variant of mediation analysis yields positive results on this dataset.

Now, the significant mediated effects above may have two distinct causal interpretations. To afford evidence in favor or against the “native” causal claim of brain-behavior mediation analysis, we derived, for each participant and each significant cluster, our I/O test statistics. Note that, prior to the analysis, we summarized the trial-by-trial variance in each cluster using the 20 first principal components from the within-cluster PCA decomposition (on average cross clusters and participants, these preserve 89% ±2% of the trial-by-trial variance). The group-level empirical distribution of $\bar{\lambda}$ is shown on the lower panel of Figure 10, for each of the 6 significant clusters. Reassuringly, all clusters exhibit strong evidence in favor of the “native” causal interpretation of brain-behavior mediation analysis, i.e. $\bar{\lambda} < 0$ for all subjects and all clusters. We will comment on these results in the Discussion section.

Now, the effect of experimental factors seems to be mediated by a set of anatomically segregated regions in the brain. These regions are likely to be organized into a functional network (cf. Figure 1 above), eventually exerting competitive and/or cooperative influences on behavioral responses. The analysis above is agnostic about the functional architecture of this network. However, the extent to which each of these network nodes actually mediates the effect of gains and losses onto choices varies across subjects. Thus, a given individual may have an idiosyncratic structure of brain pathways for processing gain and loss information. In turn, inter-individual differences in the pattern of mediated effect sizes may have behavioral consequences in terms of how strongly gains and/or losses impact decisions under risk.

Recall that the balance between the behavioral effects of gains and losses is measured using the loss aversion index $\omega = \log\left(\hat{c}_L / \hat{c}_G\right)$ (cf. Figure 9). One may thus ask whether the pattern of mediated effect sizes predicts loss aversion. We thus extracted, in each voxel of the 6 significant mediating clusters above, the indirect effect size $\left|\hat{a}_G \hat{b}\right|$, and average these within each cluster. This resulted in 6 region-specific indirect sizes per participant. We then regressed loss aversion indices against (log-transformed) indirect effect sizes, across participants. The results of this analysis are summarized on Figure 11 below.

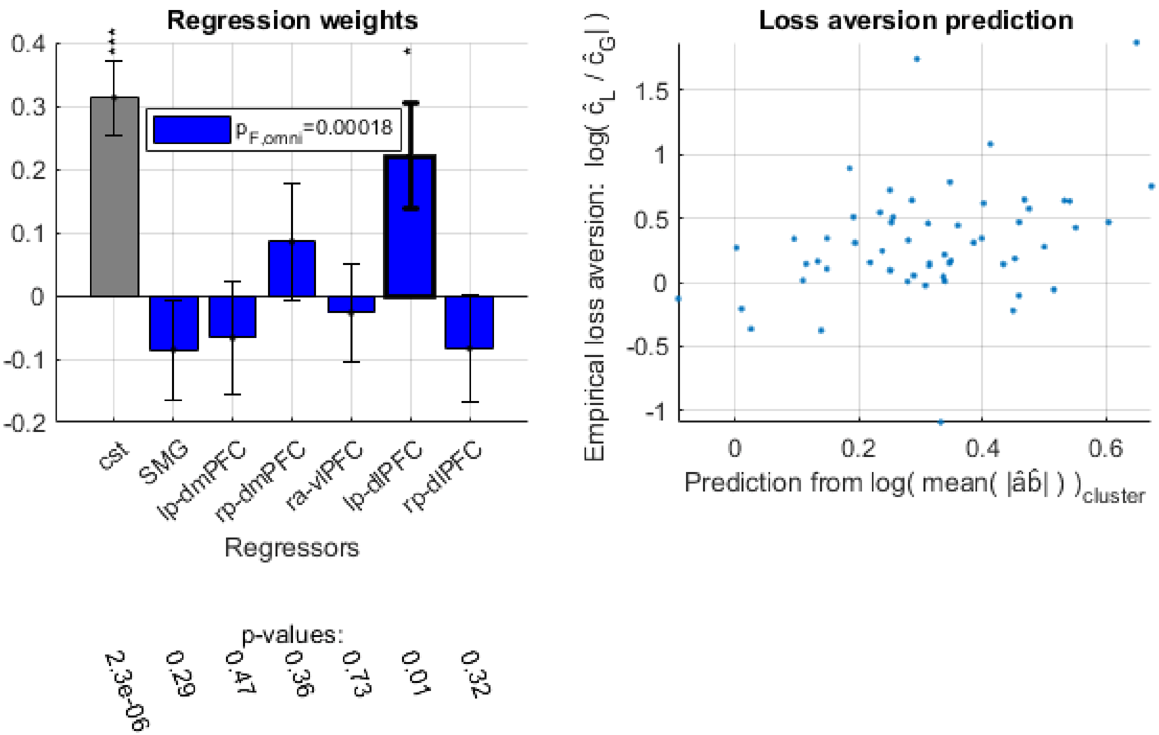

Figure 11: Inter-individual differences in loss-aversion.
Left panel: regression coefficients of the analysis of inter-subject differences of loss aversion (grey: constant term, blue: weight of inter-individual differences in cluster-averages of indirect

effect size). Errorbars depict standard errors of the mean. Right panel: Observed (y-axis) versus predicted (x-axis) loss aversion. Each dot is a participant.

First, an omnibus F-test shows that the pattern of indirect effect sizes significantly predicts loss aversion (F=5.13, dof=[7,51], R^2=12.8%, p=2x10^{-4}). This is important, since this means that one can think of loss aversion in terms of a trait that is partly determined by the relative contribution of processing pathways that mediate the effect of gains onto decisions under risk. In addition, one can see that loss aversion increases when the indirect effect size in the left posterior dmPFC increases (t=2.66, dof=51, p=0.01). No indirect effect size in any other region has a significant effect on loss aversion (all p>0.29).

Discussion

In this work, we identified the specific challenges of brain-behavior mediation analysis. In particular, we evaluated the specificity and sensitivity of five statistical tests, including so-called *indirect* and *conjunctive* approach. In brief, the *conjunctive* approach systematically shows higher sensitivity, while yielding valid inference. In addition, we disclosed the non-trivial impact of neural noise, and assessed the robustness to deviations from fMRI modelling assumptions. The former implies that brain-behavior mediation analysis cannot detect mediators that are too close from either end of the neural information processing hierarchy. *In-silico* investigations of the latter eventually favor the *convolution* approach to HRF modelling. We also disclosed some interpretational issues of mediated effects, in particular: significant mediated effects have two distinct causal interpretations. Importantly, this causal degeneracy may be partially addressed using complementary multivariate I/O test statistics. In addition, it has unexpected favorable computational consequences for whole-brain mediation analysis. Lastly, brain-behavior mediation analysis of fMRI data acquired in the context of decisions under risk further demonstrated the importance of methodological choices regarding brain-behavior mediation analysis. Eventually, the *conjunctive*/*convolution* test approach showed that the right SMG, bilateral posterior dmPFC, right anterior vlPFC and bilateral posterior dlPFC mediate the effect of prospective gains on decisions under risk. Group-level I/O test statistics provided evidence that these regions are contributing to shaping behavioral responses (in a feedforward, causal, manner), rather than collecting and/or processing information about it (cf. interpretational issue). Finally, we showed that inter-individual differences in loss aversion is partly determined by the relative contribution of these six regions to behavioral control.

Taken together, our numerical simulations and analyses of experimental fMRI data demonstrated that *conjunctive* testing has higher statistical sensitivity than *indirect* approaches. This is true even for the bias-corrected M3 bootstrap test, despite its huge computational cost. We note that the sensitivity of the M3 bootstrap test may, in principle, be improved by increasing the number of permutations used to approximate the null distribution (here: 1,000). This however, would render whole-brain analysis excessively slow. Note that M3 bootstrap and conjunctive tests had already been compared at the standard 5% significance threshold outside the context of fMRI (Hayes and Scharkow, 2013). Although authors noted that bias-corrected bootstrap tests were slightly invalid (false positive rate greater than 5%), they recommended them because they eventually yielded more reliable confidence interval estimations. We extended these simulations, eventually showing that the invalidity of bias-corrected bootstrap tests increases as one relies on more stringent significance threshold (cf. Figure 3), which is required when correcting for multiple comparisons. For all these reasons (test validity, statistical sensitivity and computational cost), we would rather favor conjunctive testing for mass-univariate brain-behavior mediation analysis.

Although computationally expedient, mass-univariate brain-behavior mediation analysis essentially relies upon an incomplete model. Not only is it agnostic about the structure of the distributed brain system that process the incoming information (cf. Figure 1), but local, voxel-based, mediation tests simply ignore about 99.999% of the brain. We would argue however, that such incompleteness may be *necessary* for statistical mediation analysis. Recall that evidence for a mediated effect requires an appropriate amount of neural noise. But neural noise estimates have two entirely distinct sources. On the one hand, it may derive from irreducible variations in neural responses that are inherent to the underlying neurobiological processes. On the other hand, it may arise from imperfections in the way neural responses are modeled. The latter most likely applies to the linear brain-behavior model in Equation 2. For example, saturating neural responses to stimuli would, under Equation 2, inflate model residuals. However, although the ensuing neural noise estimates $\hat{\varepsilon}_M^0$ would be partly artefactual, they would still be very informative to predict behavioral responses Y above and beyond the *linear* effect of X. Now, let us assume that a neurocognitive model was available, that would describe how incoming information X would be distorted, transformed and integrated with other (potentially incidental) processes, along the processing hierarchy. For example, such model may derive from recent work in theoretical neuroscience regarding population coding (Averbeck et al., 2006; Georgopoulos et al., 1986), predictive coding (Bastos et al., 2012; Hosoya et al., 2005; Rao and Ballard, 1999) or efficient coding (Barlow, 1961; Doi and Lewicki, 2011). Or it could rely on agnostic multivariate and/or nonlinear decompositions that, when properly parameterized, would account for all sorts of complex relationships between X and M. In any case, if the model was complete enough, then observed neural activity would not strongly deviate from its predictions. This would preclude the statistical detection of mediated effects. Ironically speaking then, progress in modelling neural information processing may eventually hinder the statistical efficiency of brain-behavior mediation analysis. More practically, this means that statistical brain-mediation analysis may be used in an exploratory manner, to identify brain regions that contribute to behavioral control. Further, complementary, model-based approaches to neural information processing would then help reducing one's epistemic uncertainty regarding neural noise. For example, artificial neural network modelling may be useful to identify either the structure of processing pathways (Rigoux and Daunizeau, 2015) and/or the impact of incidental biological constraints that may distort local neural information processing (Brochard and Daunizeau, 2020).

This is not to say, however, that statistical brain-behavior mediation analysis cannot be improved.

For example, one may aim at providing more informative inferences regarding the structure of the underlying processing hierarchy. A possibility here is to merge mediation analysis with existing graph analysis techniques that were developed for assessing effective connectivity in the brain (Alstott et al., 2009; Smith et al., 2011; Sporns, 2013). Another, less exhaustive but simpler, solution is to work iteratively: having identified a brain region that significantly mediates the $X \to Y$ effect, one may then look for other brain regions that would mediate both $X \to M$ and $M \to Y$ relationships, and repeat on subsequent mediators. Note that this would require additional corrections for the natural dependencies between brain regions. We refer the interested reader to Van Kesteren & Oberski (2019) for an interesting first step in this direction.

We also think that progress can be made regarding the main interpretational issue of brain-mediation analysis. In this context, let us highlight two extensions of linear mass-univariate approaches that sound promising.

First, one may rely on more stringent inferences regarding the causality of the $M \to Y$ relationship (Preacher, 2015). For example, temporal precedence may be accounted for, and inserted in the brain-mediation model using variants of Granger causality (Zhao and Luo, 2017). Note that special care must be taken regarding hemodynamic delays, whose variations across brain regions may confound temporal precedence. In particular, established fMRI applications of Granger causality are known to be prone to such confounds (David et al., 2008; Deshpande et al., 2010; Zhao and Luo, 2017). Nevertheless, constraining the brain-behavior mediation model with temporal precedence would likely reduce spurious inferences.

Second, one may exploit locally multivariate information to discriminate between many-to-one ($M \to Y$) and one-to-many ($Y \to M$) input/output mappings. In this work, we proposed a first step in this direction: namely, the I/O test statistics $\bar{\lambda}$. Numerical simulations demonstrated the utility of this information-theoretic measure, and its robustness to partial information losses that result from data dimensionality reduction. However, this work falls short of an exhaustive analytical treatment of I/O test statistics. For example, neither did we investigate whether and how nonlinearities in causal relationships confound and/or bias $\bar{\lambda}$ estimates, nor did we derive a formal statistical test of the significance of $\bar{\lambda}$ estimates. We note that the difficulty here, is that the null hypothesis may not be the most useful reference point for I/O test statistics. Rather, one aims at comparing two alternative non-nested models. Therefore, an optimal statistical treatment of I/O tests statistics may be best approached using a bayesian approach (Kass and Raftery, 1995; Liu and Aitkin, 2008). We will pursue this and related extensions of I/O test statistics in subsequent publications.

Finally, let us discuss the results of our fMRI analysis. Recall that we identified six candidate mediators of the effect of gain onto decisions under risk. Among these, the posterior dmPFC was previously shown to regulate speed-accuracy tradeoffs (Forstmann et al., 2008) and its anatomical lesion is known to impair inhibitory control in the presence of response conflict (Nachev et al., 2007). Also, decades of neuroimaging, stimulation and lesions studies have evidenced the role of posterior dlPFC and vlPFC cortices in cognitive control (Gbadeyan et al., 2016; Levy and Wagner, 2011; MacDonald et al., 2000; Miller and Cohen, 2001; Nee and D'Esposito, 2017; Soutschek and Tobler, 2020). In addition, functional and anatomical studies report convergent evidence that the right SMG is crucial for regulating emotional responses (Adolphs, 2002; Makovac et al., 2016;

Silani et al., 2013). Now, in the context of decisions under risk, automatic fear responses may induce a default tendency to reject risky gambles, eventually yielding loss aversion (Martino et al., 2006, 2010). This default emotional bias may enter in conflict with the appetitive effect of prospective gains. Whether the appetitive dimension of gambles eventually dominates automatic fear responses may then depend on the potentiation of emotional responses and on the efficiency of downstream cognitive control, which would explain why the SMG, dmPFC, dlPFC and vlPFC cortices mediate the effect of gain on decisions. This is also in line with our analysis of inter-individual differences of loss aversion, which shows that peoples' loss aversion increases when the indirect effect size (of gains on accept decisions) in the left dlPFC pathway increases. This is because a strong involvement of the dlPFC pathway may signal inefficient cognitive control (Braver et al., 2010; Poldrack, 2015), which would result in loss aversion worsening.

Although quite self-consistent and elegant, this interpretation really relies on the "native" causal interpretation of brain-behavior mediation analysis. So what if we had not found support for this causal scenario with our I/O test statistics? In fact, the existing literature may also be queried to find past evidence that may be more compatible with the alternative causal interpretation of brain-behavior mediation analysis. For example, beyond its well-known implication in language processing, the right SMG has been shown to be involved in somatosensory perception (Ben-Shabat et al., 2015; Tunik et al., 2008). Under this perspective, evidence for $b \neq 0$ (or, equivalently, $d \neq 0$) may be interpreted in terms of low-level perceptual representations of (motor?) action plans. We note that the experimental design is compatible with this interpretation because the spatial arrangement of accept/reject responses is not randomized over trials (Chen, 2014). This highlights the need for developing approaches that reduce the causal ambiguity of simple brain-behavior mediation analyses.

References

Adolphs, R. (2002). Neural systems for recognizing emotion. Curr. Opin. Neurobiol. *12*, 169–177.

Aerts, H., Fias, W., Caeyenberghs, K., and Marinazzo, D. (2016). Brain networks under attack: robustness properties and the impact of lesions. Brain J. Neurol.

Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., and Sporns, O. (2009). Modeling the Impact of Lesions in the Human Brain. PLoS Comput. Biol. *5*, e1000408.

Aroian, L.A. (1947). The Probability Function of the Product of Two Normally Distributed Variables. Ann. Math. Stat. *18*, 265–271.

Atlas, L.Y., Bolger, N., Lindquist, M.A., and Wager, T.D. (2010). Brain mediators of predictive cue effects on perceived pain. J. Neurosci. Off. J. Soc. Neurosci. *30*, 12964–12977.

Atlas, L.Y., Lindquist, M.A., Bolger, N., and Wager, T.D. (2014). Brain mediators of the effects of noxious heat on pain. Pain *155*, 1632–1648.

Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nat. Rev. Neurosci. *7*, 358–366.

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. Sens. Commun. 217–234.

Baron, R.M., and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. *51*, 1173–1182.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical Microcircuits for Predictive Coding. Neuron *76*, 695–711.

Bays, P.M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. J. Neurosci. *34*, 3632–3645.

Ben-Shabat, E., Matyas, T.A., Pell, G.S., Brodtmann, A., and Carey, L.M. (2015). The Right Supramarginal Gyrus Is Important for Proprioception in Healthy and Stroke-Affected Participants: A Functional MRI Study. Front. Neurol. *6*.

Braver, T.S., Cole, M.W., and Yarkoni, T. (2010). Vive les differences! Individual variation in neural mechanisms of executive control. Curr. Opin. Neurobiol. *20*, 242–250.

Brochard, J., and Daunizeau, J. (2020). Blaming blunders on the brain: can indifferent choices be driven by range adaptation or synaptic plasticity? BioRxiv 2020.09.08.287714.

Chen, M.-Y. (2014). The development of bias in perceptual and financial decision-making. Thesis.

Chén, O.Y., Crainiceanu, C., Ogburn, E.L., Caffo, B.S., Wager, T.D., and Lindquist, M.A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. Biostat. Oxf. Engl. *19*, 121–136.

Csató, L., and Opper, M. (2003). Sparse gaussian processes: inference, subspace identification and model selection. IFAC Proc. Vol. *36*, 789–794.

David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., and Depaulis, A. (2008). Identifying Neural Drivers with Functional MRI: An Electrophysiological Validation. PLoS Biol *6*, e315.

Deshpande, G., Sathian, K., and Hu, X. (2010). Effect of hemodynamic variability on Granger causality analysis of fMRI. NeuroImage *52*, 884–896.

Dinstein, I., Heeger, D.J., and Behrmann, M. (2015). Neural variability: friend or foe? Trends Cogn. Sci. *19*, 322–328.

Doi, E., and Lewicki, M.S. (2011). Characterization of Minimum Error Linear Coding with Sensory and Neural Noise. Neural Comput. *23*, 2498–2510.

Faisal, A.A., Selen, L.P.J., and Wolpert, D.M. (2008). Noise in the nervous system. Nat. Rev. Neurosci. *9*, 292–303.

Ferster, D. (1996). Is Neural Noise Just a Nuisance? Science *273*, 1812–1812.

Flandin, G., and Friston, K.J. (2019). Analysis of family-wise error rates in statistical parametric mapping using random field theory. Hum. Brain Mapp. *40*, 2052–2054.

Forstmann, B.U., Dutilh, G., Brown, S., Neumann, J., Cramon, D.Y. von, Ridderinkhof, K.R., and Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. Proc. Natl. Acad. Sci. *105*, 17538–17542.

Friston, K.J. (2011). Functional and Effective Connectivity: A Review. Brain Connect. *1*, 13–36.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C., Frackowiak, R.S., and Turner, R. (1995). Analysis of fMRI time-series revisited. NeuroImage *2*, 45–53.

Friston, K.J., Holmes, A.P., Price, C.J., Büchel, C., and Worsley, K.J. (1999). Multisubject fMRI Studies and Conjunction Analyses. NeuroImage *10*, 385–396.

Friston, K.J., Penny, W.D., and Glaser, D.E. (2005a). Conjunction revisited. NeuroImage *25*, 661–667.

Friston, K.J., Stephan, K.E., Lund, T.E., Morcom, A., and Kiebel, S. (2005b). Mixed-effects and fMRI studies. NeuroImage *24*, 244–252.

Gbadeyan, O., McMahon, K., Steinhauser, M., and Meinzer, M. (2016). Stimulation of Dorsolateral Prefrontal Cortex Enhances Adaptive Cognitive Control: A High-Definition Transcranial Direct Current Stimulation Study. J. Neurosci. *36*, 12530–12536.

Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1986). Neuronal population coding of movement direction. Science *233*, 1416–1419.

Geuter, S., Losin, E.A.R., Roy, M., Atlas, L.Y., Schmidt, L., Krishnan, A., Koban, L., Wager, T.D., and Lindquist, M.A. (2018). Multiple brain networks mediating stimulus-pain relationships in humans. BioRxiv 298927.

Gitelman, D.R., Penny, W.D., Ashburner, J., and Friston, K.J. (2003). Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. NeuroImage *19*, 200–207.

Goodman, L.A. (1960). On the Exact Variance of Products. J. Am. Stat. Assoc. *55*, 708–713.

Hayes, A.F., and Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: does method really matter? Psychol. Sci. *24*, 1918–1927.

He, Y., and Evans, A. (2010). Graph theoretical modeling of brain connectivity. Curr. Opin. Neurol. *23*, 341–350.

Holmes, A.P., Friston, K.J., and Friston, K. (1998). Generalisability, random effects and population inference.

Hong, S.L., and Rebec, G.V. (2012). A new perspective on behavioral inconsistency and neural noise in aging: compensatory speeding of neural communication. Front. Aging Neurosci. *4*.

Hosoya, T., Baccus, S.A., and Meister, M. (2005). Dynamic predictive coding by the retina. Nature *436*, 71–77.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. J. Am. Stat. Assoc. *90*, 773–795.

Kenny, D.A., Korchmaros, J.D., and Bolger, N. (2003). Lower level mediation in multilevel models. Psychol. Methods *8*, 115–128.

Koban, L., Kross, E., Woo, C.-W., Ruzic, L., and Wager, T.D. (2017). Frontal-Brainstem Pathways Mediating Placebo Effects on Social Rejection. J. Neurosci. *37*, 3621–3631.

Koban, L., Jepma, M., López-Solà, M., and Wager, T.D. (2019). Different brain networks mediate the effects of social and conditioned expectations on pain. Nat. Commun. *10*, 4096.

Levy, B.J., and Wagner, A.D. (2011). Cognitive control and right ventrolateral prefrontal cortex: reflexive reorienting, motor inhibition, and action updating. Ann. N. Y. Acad. Sci. *1224*, 40–62.

Liao, C.H., Worsley, K.J., Poline, J.-B., Aston, J.A.D., Duncan, G.H., and Evans, A.C. (2002). Estimating the Delay of the fMRI Response. NeuroImage *16*, 593–606.

Lindquist, M.A. (2012). Functional causal mediation analysis with an application to brain connectivity. J. Am. Stat. Assoc. *107*, 1297–1309.

Lindquist, M.A., and Mejia, A. (2015). Zen and the Art of Multiple Comparisons. Psychosom. Med. *77*, 114.

Liu, C.C., and Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. J. Math. Psychol. *52*, 362–375.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature *412*, 150–157.

MacDonald, A.W., Cohen, J.D., Stenger, V.A., and Carter, C.S. (2000). Dissociating the Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control. Science *288*, 1835–1838.

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. Psychol. Methods *7*, 83–104.

MacKinnon, D.P., Lockwood, C.M., and Williams, J. (2004). Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. Multivar. Behav. Res. *39*, 99.

MacKinnon, D.P., Fairchild, A.J., and Fritz, M.S. (2007). Mediation Analysis. Annu. Rev. Psychol. *58*, 593.

Makovac, E., Meeten, F., Watson, D.R., Garfinkel, S.N., Critchley, H.D., and Ottaviani, C. (2016). Neurostructural abnormalities associated with axes of emotion dysregulation in generalized anxiety. NeuroImage Clin. *10*, 172–181.

Marrelec, G., Daunizeau, J., Pelegrini-Issac, M., Doyon, J., and Benali, H. (2005). Conditional correlation as a measure of mediated interactivity in fMRI and MEG/EEG. IEEE Trans. Signal Process. *53*, 3503–3516.

Martin, C., Martindale, J., Berwick, J., and Mayhew, J. (2006). Investigating neural-hemodynamic coupling and the hemodynamic response function in the awake rat. NeuroImage *32*, 33–48.

Martino, B.D., Kumaran, D., Seymour, B., and Dolan, R.J. (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. Science *313*, 684–687.

Martino, B.D., Camerer, C.F., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. Proc. Natl. Acad. Sci. *107*, 3788–3792.

McDonnell, M.D., and Ward, L.M. (2011). The benefits of noise in neural systems: bridging theory and experiment. Nat. Rev. Neurosci. *12*, 415–426.

McGill, W.J. (1954). Multivariate information transmission. Psychometrika *19*, 97–116.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Moran, P. a. P. (1970). On asymptotically optimal tests of composite hypotheses. Biometrika *57*, 47–55.

Nachev, P., Wydell, H., O'Neill, K., Husain, M., and Kennard, C. (2007). The role of the pre-supplementary motor area in the control of action. Neuroimage *36*, T155–T163.

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. NeuroImage *56*, 400–410.

Nee, D.E., and D'Esposito, M. (2017). Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. ELife *6*, e28040.

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. NeuroImage *25*, 653–660.

Palestro, J.J., Bahg, G., Sederberg, P.B., Lu, Z.-L., Steyvers, M., and Turner, B.M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. J. Math. Psychol. *84*, 20–48.

Pearl, J. (2012). The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models. In Causality, (John Wiley & Sons, Ltd), pp. 151–179.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., and Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. NeuroImage *104*, 209–220.

Poldrack, R.A. (2015). Is "efficiency" a useful concept in cognitive neuroscience? Dev. Cogn. Neurosci. *11*, 12–17.

Poldrack, R.A., Barch, D.M., Mitchell, J., Wager, T., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., and Milham, M. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. Front. Neuroinformatics *7*.

Preacher, K.J. (2015). Advances in Mediation Analysis: A Survey and Synthesis of New Developments. Annu. Rev. Psychol. *66*, 825–852.

Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

Rigoux, L., and Daunizeau, J. (2015). Dynamic causal modelling of brain-behaviour relationships. NeuroImage *117*, 202–221.

Robbins, T.W. (2011). Cognition: The Ultimate Brain Function. Neuropsychopharmacology *36*, 1–2.

Rosenblatt, M. (1956). A CENTRAL LIMIT THEOREM AND A STRONG MIXING CONDITION. Proc. Natl. Acad. Sci. U. S. A. *42*, 43–47.

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. NeuroImage *52*, 1059–1069.

Seymour, B., O'Doherty, J.P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat. Neurosci. *8*, 1234–1240.

Shadlen, M.N., and Newsome, W.T. (1994). Noise, neural codes and cortical organization. Curr. Opin. Neurobiol. *4*, 569–579.

Silani, G., Lamm, C., Ruff, C.C., and Singer, T. (2013). Right Supramarginal Gyrus Is Crucial to Overcome Emotional Egocentricity Bias in Social Judgments. J. Neurosci. *33*, 15466–15476.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., and Woolrich, M.W. (2011). Network modelling methods for FMRI. NeuroImage *54*, 875–891.

Sobel, M.E. (Ed) (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociol Methodol *13*, 290–312.

Soutschek, A., and Tobler, P.N. (2020). Causal role of lateral prefrontal cortex in mental effort and fatigue. Hum. Brain Mapp.

Sporns, O. (2013). Making sense of brain network data. Nat. Methods *10*, 491–493.

Stein, R.B., Gossen, E.R., and Jones, K.E. (2005). Neuronal variability: noise or part of the signal? Nat. Rev. Neurosci. *6*, 389–397.

Tunik, E., Lo, O.-Y., and Adamovich, S.V. (2008). Transcranial Magnetic Stimulation to the Frontal Operculum and Supramarginal Gyrus Disrupts Planning of Outcome-Based Hand–Object Interactions. J. Neurosci. *28*, 14422–14427.

Turner, B.M., Palestro, J.J., Miletić, S., and Forstmann, B.U. (2019). Advances in techniques for imposing reciprocity in brain-behavior relations. Neurosci. Biobehav. Rev. *102*, 327–336.

Wager, T. (2008). canlab/ M3 MediationToolbox (Cognitive and Affective Neuroscience Laboratory).

Wager, T.D., Davidson, M.L., Hughes, B.L., Lindquist, M.A., and Ochsner, K.N. (2008). Prefrontal-Subcortical Pathways Mediating Successful Emotion Regulation. Neuron *59*, 1037–1050.

Wager, T.D., Waugh, C.E., Lindquist, M., Noll, D.C., Fredrickson, B.L., and Taylor, S.F. (2009a). Brain mediators of cardiovascular responses to social threat: part I: Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. NeuroImage *47*, 821–835.

Wager, T.D., van Ast, V.A., Hughes, B.L., Davidson, M.L., Lindquist, M.A., and Ochsner, K.N. (2009b). Brain mediators of cardiovascular responses to social threat, part II: Prefrontal-subcortical pathways and relationship with anxiety. NeuroImage *47*, 836–851.

Woo, C.-W., Roy, M., Buhle, J.T., and Wager, T.D. (2015). Distinct Brain Systems Mediate the Effects of Nociceptive Input and Self-Regulation on Pain. PLOS Biol. *13*, e1002036.

Worsley, K.J., and Friston, K.J. (1995). Analysis of fMRI time-series revisited--again. NeuroImage *2*, 173–181.

Yamamoto, D.J., Woo, C.-W., Wager, T.D., Regner, M.F., and Tanabe, J. (2015). Influence of dorsolateral prefrontal cortex and ventral striatum on risk avoidance in addiction: a mediation analysis. Drug Alcohol Depend. *149*, 10–17.

Zhang, S., Mano, H., Lee, M., Yoshida, W., Robbins, T., Kawato, M., and Seymour, B. (2017). The Control of Tonic Pain by Active Relief Learning. BioRxiv 222653.

Zhao, Y., and Luo, X. (2017). Granger Mediation Analysis of Multiple Time Series with an Application to fMRI. ArXiv170905328 Stat.

(2010). Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control (New York: Springer-Verlag).

Appendix A: OLS estimators of path coefficients

Recall that the second line of Equation 2 can be re-written as:

$$\begin{aligned} Y &= Mb + Xc' + \varepsilon_Y \\ &= \begin{bmatrix} M & X \end{bmatrix} \begin{bmatrix} b \\ c' \end{bmatrix} + \varepsilon_Y \end{aligned} \tag{A1}$$

The OLS estimators of b and c' path coefficients are thus given by:

$$\begin{aligned} \begin{bmatrix} \hat{b} \\ \hat{c}' \end{bmatrix} &= \left(\begin{bmatrix} M^T \\ X^T \end{bmatrix} \begin{bmatrix} M & X \end{bmatrix} \right)^{-1} \begin{bmatrix} M^T \\ X^T \end{bmatrix} Y \\ &= \begin{bmatrix} n & M^T X \\ X^T M & n \end{bmatrix}^{-1} \begin{bmatrix} M^T \\ X^T \end{bmatrix} Y \\ &= \frac{1}{n^2 - M^T X X^T M} \begin{bmatrix} n & -M^T X \\ -X^T M & n \end{bmatrix} \begin{bmatrix} M^T \\ X^T \end{bmatrix} Y \\ &= \frac{1}{n^2 - M^T X X^T M} \begin{bmatrix} nM^T - M^T X X^T \\ nX^T - X^T M M^T \end{bmatrix} Y \end{aligned} \tag{A2}$$

where the third line derives from the analytical formulation of 2x2 inverse matrices.

Now recall that $M = X\hat{a} + \hat{\varepsilon}_M^0$ with $\hat{a} = 1/n\, M^T X$. The estimator of path coefficients b and c' thus writes:

$$\begin{aligned} \begin{bmatrix} \hat{b} \\ \hat{c}' \end{bmatrix} &= \frac{1}{n^2 - n^2 \hat{a}^2} \begin{bmatrix} n\left(X\hat{a} + \hat{\varepsilon}_M^0 \right)^T - n\hat{a}X^T \\ nX^T - n\hat{a}\left(X\hat{a} + \hat{\varepsilon}_M^0 \right)^T \end{bmatrix} Y \\ &= \frac{1}{n - n\hat{a}^2} \begin{bmatrix} \hat{\varepsilon}_M^{0\,T} \\ \left(1 - \hat{a}^2\right) X^T - \hat{a}\hat{\varepsilon}_M^{0\,T} \end{bmatrix} Y \\ &= \begin{bmatrix} \dfrac{1}{1-\hat{a}^2} \dfrac{1}{n} \hat{\varepsilon}_M^{0\,T} Y \\ \dfrac{1}{n} X^T Y - \hat{a}\hat{b} \end{bmatrix} \end{aligned} \tag{A3}$$

This completes the derivation of path coefficients' estimates.

Appendix B: Sobel's test.

In what follows, we summarize the derivation of Sobel's mediation test.

First, recall that, given Equations 4 and 5, both $\hat{a}$ and $\hat{b}$ follow gaussian distributions, namely: $\hat{a} \sim N\left(a, \hat{\sigma}_a^2\right)$ and $\hat{b} \sim N\left(b, \hat{\sigma}_b^2\right)$. Sobel's approach effectively reduces to a Laplace approximation of the distribution of the product $\hat{a}\hat{b}$ of path coefficients' estimates.

Let $f\left(\hat{a}, \hat{b}\right) = \hat{a}\hat{b}$ be the function that maps the pair of path coefficient estimates to their product. One can approximate $f\left(\hat{a}, \hat{b}\right)$ using a first-order Taylor expansion in the neighborhood of some arbitrary point $\left(a_0, b_0\right)$:

$$\begin{aligned} f\left(\hat{a}, \hat{b}\right) &\approx f\left(a_0, b_0\right) + \left.\frac{\partial f}{\partial \hat{a}}\right|_{a_0, b_0} \left(\hat{a} - a_0\right) + \left.\frac{\partial f}{\partial \hat{b}}\right|_{a_0, b_0} \left(\hat{b} - b_0\right) \\ &= a_0 b_0 + b_0\left(\hat{a} - a_0\right) + a_0\left(\hat{b} - b_0\right) \\ &= a_0 \hat{b} + b_0 \hat{a} - a_0 b_0 \end{aligned} \qquad \text{(A4)}$$

If we choose to use the above Taylor expansion in the neighborhood of the unknown true values of path coefficients $\left(a_0 \equiv a, b_0 \equiv b\right)$, then Equation A4 provides us with a Laplace approximation to the first two moments of the bivariate product $\hat{a}\hat{b}$:

$$\begin{cases} E\left[\hat{a}\hat{b}\right] \approx ab \\ Var\left[\hat{a}\hat{b}\right] \approx \hat{\sigma}_a^2 b^2 + \hat{\sigma}_b^2 a^2 \end{cases} \qquad \text{(A5)}$$

The Sobel test directly relies on this approximation to form a pseudo z-score $z_{ab}^{(Sobel)}$ for the strength of the indirect path, as follows:

$$z_{ab}^{(Sobel)} = \frac{\hat{a}\,\hat{b}}{\sqrt{\hat{\sigma}_a^2\hat{b}^2 + \hat{\sigma}_b^2\hat{a}^2}} \tag{A6}$$

where the unknown path coefficients have been replaced by their OLS estimates. Note that $z_{ab}^{(Sobel)}$ is invariant under arbitrary rescaling of X, Y and/or M. Under the null $H_0 : ab = 0$, $z_{ab}^{(Sobel)}$ approximately follows Student's probability density function with appropriate degrees of freedom.

We note that this approximation will be quite tight away from the diagonal lines $\hat{a} = \pm\hat{b}$, where the product $\hat{a}\hat{b}$ will start to behave as a quadratic function. But Sobel's approximation error will grow quicker than estimation errors on path coefficients.

One can also show that Sobel’s test statistics is always smaller than conjunctive’s test statistics:

$$\begin{aligned}
\left|z_{ab}^{(Sobel)}\right| &= \frac{1}{\sqrt{\hat{\sigma}_a^2/\hat{a}^2 + \hat{\sigma}_b^2/\hat{b}^2}} \\
&= \frac{1}{\sqrt{1/t_a{}^2 + 1/t_b{}^2}} \\
&= \frac{|t_a||t_b|}{\sqrt{t_a{}^2 + t_b{}^2}} \\
&= \min\left(|t_a|,|t_b|\right)\frac{\max\left(|t_a|,|t_b|\right)}{\sqrt{t_a{}^2 + t_b{}^2}} \\
&\leq \min\left(|t_a|,|t_b|\right)
\end{aligned} \tag{A7}$$

where $\min\left(|t_a|,|t_b|\right)$ is the conjunctive test statistics (cf. Equation 9).

Appendix C: Dealing with contrasts on experimental conditions

So far, we have only considered simple independent variables X. However, a typical experiment includes more than one condition or factor, and the question of interest might be best framed in terms of mediating the effect of a linear combination of independent variables. In other terms, we want to generalize classical mediation analyses of the sort implied by Equation 1 to *contrasts* of experimental factors.

Without loss of generality, let us consider an experimental design with n_{cond} conditions, which are encoded through a $n \times n_{cond}$ design matrix $\mathbf{X}$. Typically, the entries of $\mathbf{X}$'s columns would be zero everywhere, except at trials that belong to the corresponding condition (where their value would be one). Replacing X with the design matrix $\mathbf{X}$ in Equation 2 induces the following two-fold lieanr regression model:

$$\begin{cases} M = \mathbf{Xa} + \varepsilon_M \\ Y = Mb + \mathbf{Xc}' + \varepsilon_Y \end{cases} \tag{A8}$$

where $\mathbf{a}$ and $\mathbf{c}'$ are now $n_{cond} \times 1$ vectors of regression coefficients that encode the condition means. In this context, most experimental questions of interest are framed in terms of contrasts on path coefficients $\mathbf{a}$. So how can one ask whether M mediates the effect of an arbitrary contrast on path coefficients?

Two cases may arise. In the simplest case, one would deal with single contrasts. Let $\mathbf{w}$ be an arbitrary $n_{cond} \times 1$ vector of contrast weights. For example, in a typical 2×2 factorial design, $\mathbf{w} = [1 \quad -1 \quad -1 \quad 1]$ would be capturing the interaction between the two factors. Single contrasts of this sort do not require any specific adaptation of mediation analyses, because $\mathbf{w}^T\mathbf{a}$ is a scalar, and its OLS estimate has a known fixed-form distribution under the null. In turn, asking whether single contrasts are mediated by M reduce to testing whether $(\mathbf{w}^T\mathbf{a})\, b \neq 0$, which can be done using either the indirect or conjunctive approaches described above. Slightly more subtle is the case of multiple contrasts, as induced by global null hypotheses tests. For example, let us consider an experimental design with three conditions. In analogy to ANOVA, we wish to test for the mediation of *any* difference between the conditions. The corresponding contrast of interest $\mathbf{w}$ is now a 3×2 matrix of weights, and $\mathbf{w}^T\mathbf{a}$ becomes a 2×1 vector. Outside the context of brain-behavior mediation analysis, assessing the statistical significance of such a contrast would be performed using an F-test, for which p-values can be derived analytically (Friston et al., 1995). When using the conjunctive approach, this poses no problem, as one would simply compute the p-value of the resulting minimum F-statistics. The indirect approach is more difficult to adapt here. In principle, one would first have to partition the design matrix $\mathbf{X} \leftarrow \left[\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_0\right]$ into subspaces respectively spanning the contrast of interest $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{w}$ and the effects of no interest $\tilde{\mathbf{X}}_0 = \mathbf{X}\left(\mathbf{I} - \mathbf{w}\,\mathbf{w}^-\right)$. Then, one would remove the effects of no interest from both the mediator and the dependant variable. Finally, one would have to test whether *any* indirect path induced by the ensuing columns of $\tilde{\mathbf{X}}$ is significant. The latter issue is not entirely trivial, but can be solved using the *minimum p-value* approach (Friston et al., 1999; Nichols et al., 2005) of conjunction analysis.

Appendix D: group-level random-effect analysis

Let us now consider the specific issue of experiments performed with multiple subjects. For example, let us assume that each subject participates in an experiment consisting of multiple trials, such that Equation 2 describes the relationship existing between X, M and Y across trials, at the subject-level. We now want to ask whether there is a mediated effect that is consistent across subjects, at the group-level. This calls for mixed-effects analyses, which essentially assume that subject-level path coefficients are sampled from a parent (population) distribution whose mean we wish to infer on. This can be efficiently performed using a summary statistics approach (Friston et al., 2005b; Holmes et al., 1998), whereby one first estimates subject-level effects (here, $\hat{a}_i$ and $\hat{b}_i$, where $i \in [1,\ldots n]$ is the participant's index), and then report these for a random-effect analysis at the group-level.

Similarly to subject-level analysis, both conjunctive and indirect approaches are possible here. Let μ_a and μ_b be the unknown population mean of a and b path coefficients, respectively. At the group-level, the null hypothesis of mediation analysis can be written as follows:

$$\begin{cases} H_0^{(conjunction)} : \mu_a = 0 \; OR \; \mu_b = 0 \\ H_0^{(indirect)} : \mu_a \mu_b = 0 \end{cases} \qquad \text{(A9)}$$

where $H_0^{(conjunction)} \Leftrightarrow H_0^{(indirect)}$ as before.

The conjunctive approach then simply reduces to testing whether both group-mean estimates $\hat{\mu}_a = 1/n \sum_i \hat{a}_i$ and $\hat{\mu}_b = 1/n \sum_i \hat{b}_i$ differ from zero, which can be tested using the p-value for the minimum statistic.

The indirect approach relies on testing whether the group-mean of the indirect effect differs from zero. According to the central limit theorem (Rosenblatt, 1956), the distribution of the average product $1/n \sum_i \hat{a}_i \hat{b}_i$ will quickly tend towards a Gaussian distribution. However, any non-zero covariance between path coefficients will bias the inference, because $E\left[\hat{a}\hat{b}\right] = E\left[\hat{a}\right] E\left[\hat{b}\right] + \mathrm{cov}\left[\hat{a}, \hat{b}\right]$ (Kenny et al., 2003). In other terms, even if the null hypothesis is true, covarying fluctuations in path estimates may significantly differ from zero. This is why the indirect approach should rather rely on testing the product $\hat{\mu}_a\hat{\mu}_b$ of group-mean estimates. This can be done using either parametric (cf. Sobel, Airoian or Goodman test statistics) or non-parametric (cf. M3 bootstrap) approaches.

Appendix E: Causal impact of neural noise

The non-trivial impact of neural noise is not a feature of univariate linear brain-behavior mediation models. In fact, one can show that this generalizes to any form of brain-behavior. In what follows, we rely on an information-theoretic framework that was developed for addressing mediation claims, irrespective of the mathematical form that the mediation model may take (Pearl, 2012). The only requirement here, is that of a causal cascade from X to M and Y, and from M to Y (cf. directed acyclic graph in Figure 2, left panel).

Let $IE_{XX'}(Y)$ be the expected impact of the mediator variable on the behavior, under a virtual change of the manipulation (from $X = x$ to $X = x'$, see Equation 9 in Pearl, 2012):

$$IE_{xx'}(Y) = \sum_{M} E\left[Y \middle| X, M\right]\left(P\left(M|x'\right) - P\left(M|x\right)\right) \tag{A10}$$

where $P\left(M|X\right)$ is the conditional distribution of the mediator variable. Note that, in Equation A10, the causal relationships between X, M and Y are implicitly absorbed in conditional distributions. In brief, $IE_{xx'}(Y)$ measures the strength of the indirect effect of X onto Y, i.e. it serves as a summary statistics for significance tests of (possible multivariate and nonlinear) mediated effects.

Now, when there is no neural noise, the mediator variable brings no additional information on the behavior, i.e. $E\left[Y|X,M\right] \approx E\left[Y|X\right]$. In turn, the mediator's impact $IE_{xx'}\left(Y\right)$ becomes negligible: $IE_{xx'}\left(Y\right) \approx E\left[Y|X\right]\left(\sum_M P\left(M|x'\right) - \sum_M P\left(M|x\right)\right) = 0$. In other terms, when the mediator brings no additional information on behavior, it cannot be detected.

Conversely, when neural noise dominates, the mediator is effectively independent from the manipulation, i.e.: $P\left(M|x'\right) \approx P\left(M|x\right) \approx P\left(M\right)$. It follows that, here again: $IE_{xx'}\left(Y\right) \approx \sum_M E\left[Y|X,M\right]\left(P\left(M\right) - P\left(M\right)\right) = 0$. In other words, when the manipulation brings no or little information on the mediator, no mediation can be detected.

In conclusion, mediated effects can only be detected for intermediate neural noise magnitudes, irrespective of the mathematical form of the brain-behavior mediation model.

Appendix F: Equivalence of causal interpretations of mediation analysis

In what follows we give a proof of (i) Equation 12 in the main text, and (ii) equality of t-statistics of "native" and "swapped" path coefficients.

First one can use the expressions of their OLS estimates to derive the ratio of the two path coefficients (see Appendix A):

$$\begin{aligned}\frac{\hat{b}}{\hat{d}} &= \frac{\left(nM^T - M^T XX^T\right)Y}{n^2 - M^T XX^T M} \times \frac{n^2 - Y^T XX^T Y}{\left(nY^T - Y^T XX^T\right)M} \\ &= \frac{1-\left(Y^T X/n\right)^2}{1-\left(M^T X/n\right)^2} \times \frac{\left(nM^T - M^T XX^T\right)Y}{\left(nY^T - Y^T XX^T\right)M} \\ &= \frac{1-\hat{\rho}\left(X,Y\right)^2}{1-\hat{\rho}\left(X,M\right)^2} \times \frac{M^T Y/n - M^T X/n \times Y^T X/n}{Y^T M/n - Y^T X/n \times X^T M/n} \\ &= \frac{1-\hat{\rho}\left(X,Y\right)^2}{1-\hat{\rho}\left(X,M\right)^2}\end{aligned} \qquad \text{(A11)}$$

where $\hat{\rho}(\cdot,\cdot)$ is the sample correlation between arbitrary vectors.

Now, recall that, in Equations 1-2, (i) all mediation variables are z-scored and (ii) residual estimates are, by construction, orthogonal to the variable X . Therefore :

$$\begin{cases} 1 = \hat{\rho}(X,Y)^2 + \hat{\sigma}_{Y|X}{}^2 \\ 1 = \hat{\rho}(X,M)^2 + \hat{\sigma}_{M|X}{}^2 \end{cases} \qquad \text{(A12)}$$

This concludes the demonstration of Equation 12 of the main text ($\hat{b} = \hat{d} \times \hat{\sigma}^2_{Y|X} \,/\, \hat{\sigma}^2_{M|X}$).

Now let us prove the equality of t-statistics of "native" and "swapped" path coefficients. Using the definition of these test statistics we have:

$$\begin{aligned} & t_b = t_d \\ \Leftrightarrow\ & \frac{\hat{b}}{\sigma^2_{Y|X,M}} \times \sqrt{n-2} = \frac{\hat{d}}{\sigma^2_{M|X,Y}} \times \sqrt{n-2} \\ \Leftrightarrow\ & \frac{\hat{b}}{\hat{d}} = \frac{\hat{\sigma}^2_{Y|X,M}}{\hat{\sigma}^2_{M|X,Y}} \\ \Leftrightarrow\ & \frac{\hat{\sigma}^2_{Y|X}}{\hat{\sigma}^2_{M|Y}} = \frac{\hat{\sigma}^2_{Y|X,M}}{\hat{\sigma}^2_{M|X,Y}} \end{aligned} \qquad \text{(A13)}$$

Now recall the iterative decomposition of the determinant of a gram matrix: $\det\left(\left[A,v\right]^T\left[A,v\right]\right)=\det\left(A^TA\right)\times\left(v^Tv-v^TA\left(A^TA\right)^{-1}A^Tv\right)$, where A and v are arbitrary matrix and vectors, respectively (Csató and Opper, 2003). This yields:

$$\begin{aligned}\det\left(\left[X,M,Y\right]^T\left[X,M,Y\right]\right)&=n\times\det\left(\left[X,M\right]^T\left[X,M\right]\right)\times\hat{\sigma}^2_{Y|X,M}\\&=n^2\times\det\left(X^TX\right)\times\hat{\sigma}^2_{M|X}\hat{\sigma}^2_{Y|X,M}\end{aligned}\qquad\text{(A14)}$$

Similarly, we have:

$$\det\left(\left[X,Y,M\right]^T\left[X,Y,M\right]\right)=n^2\times\det\left(X^TX\right)\times\hat{\sigma}^2_{Y|X}\hat{\sigma}^2_{M|X,Y}\qquad\text{(A15)}$$

Lastly, because the order of the matrix's columns leaves the determinant unchanged, Equations A14 and A15 are identical. This implies that $\hat{\sigma}^2_{Y|X}\hat{\sigma}^2_{M|X,Y}=\hat{\sigma}^2_{M|X}\hat{\sigma}^2_{Y|X,M}$, which concludes our proof.