nature

https://doi.org/10.1038/s41586-020-2314-9

Supplementary information

Variability in the analysis of a single neuroimaging dataset by many teams

In the format provided by the authors and unedited

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols ✉, Russell A. Poldrack ✉ & Tom Schonberg ✉

Supplementary Information

Variability in the analysis of a single neuroimaging dataset by many teams

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M.W.J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W Cox, William A Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y.C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson , Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S.L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols*, Russell A. Poldrack*, Tom Schonberg*

* Corresponding author

Supplementary Methods and Results

Results of analysis teams

We performed exploratory analyses to test whether the teams' confidence ("How confident are you about this result?") and similarity ("How similar do you think your result is to the other analysis teams?") ratings were related to their final reported results. First, we ran a mixed effects logistic regression with the binary reported result as the outcome variable and the confidence and similarity ratings as predictors (the hypothesis number was also added to the model as a factorial fixed effect). We found that higher similarity rating was associated with higher proportion of significant results ($p < 0.001$, delta pseudo-$R^2 = 0.032$, mean similarity rating: 7.40 for significant results and 6.29 for insignificant results). Confidence ratings were not significantly related to the proportion of significant results ($p = 0.732$, mean confidence rating: 7.59 for significant results and 6.93 for insignificant results). Second, we tested the Spearman correlation of the distance of the outcome (the proportion of teams that reported a significant result for each hypothesis) from 0.5 (i.e., how consistent the results were across teams) and the mean confidence rating across hypotheses. This correlation was positive ($r = 0.69$, $p = 0.039$), indicating that for hypotheses where variability of the results across teams was smaller, the teams were more confident in their results. The Spearman correlation between the distance of the outcome from 0.5 and the mean estimated similarity to other teams was not significant ($r = 0.40$, $p = 0.286$).

Variability of thresholded statistical maps

We performed a coordinate-based meta-analysis using activation likelihood estimation (ALE)[30,31] across teams. This analysis, which imposes additional smoothing, was performed with the NIMARE software package [RRID:SCR_017398] using peak locations identified from thresholded maps for each team. Correction for multiple tests was applied using false discovery rate at the 5% threshold[41]. The ALE analysis demonstrated convergent patterns of activation for all hypotheses. However, while ALE has been shown to be robust to correlated inputs, in the case when some studies contribute multiple contrasts[42], the present single-study same-data usage goes beyond existing research and the extent of any potential biases is unknown. Therefore, this analysis provides only a qualitative aggregation and cannot be regarded as a calibrated statistical result due to the single-study same-data usage.

Variability of unthresholded statistical maps

Correlations between unthresholded maps were further assessed by modeling the median Spearman correlation of each team with the average pattern across teams as a function of analysis method using linear regression. Estimated spatial smoothness of the statistical images (averaged across hypotheses) was significantly associated with correlation with the mean pattern ($p = 0.023$, delta $r^2 = 0.07$), as was the use of movement modeling ($p = 0.021$, delta $r^2 = 0.08$).

No teams were consistently anticorrelated with the mean pattern across all hypotheses, though three teams showed a correlation of $r < 0.2$ with the mean pattern across hypotheses, whereas 32 teams showed correlations of $r > 0.7$ with the mean pattern.

We further performed an image-based meta-analysis (IBMA) to quantify the evidence for each hypothesis across analysis teams, accounting for the lack of independence due to the use of a common dataset across teams (Extended Data Figure 3b). While there are different meta-analysis-inspired approaches that could be taken (e.g. a random effects meta-analysis that penalizes for inter-team variation), we sought an approach that would preserve the typical characteristics of the teams' maps. In particular, the meta-analytical statistical map is based on the mean of teams' statistical maps, but is shifted and scaled by global factors so that the mean and variance are equal to the original image-wise means and variances averaged over teams. Under a complete null hypothesis of no signal anywhere for every team and every voxel, the resulting map can be expected to produce nominal standard normal z-scores, and in the presence of signal will reflect a consensus of the different results.

The image-based meta-analysis method is as follows. Let N be the number of teams, μ be the (scalar) mean over space of each team's map, averaged over teams, σ^2 likewise the spatial variance averaged over teams, and let $\mathbf{Q}$ be the $N \times N$ correlation matrix, computed using all voxels in the statistical map. Then let Z_{ik} be the z-value for voxel i and team k, and M_i the mean of those N z-values at voxel i. The variance of M_i is $\sigma^2 \mathbf{1}^\top \mathbf{Q} \mathbf{1}/N^2$, where $\mathbf{1}$ is a N-vector of ones. We center and standardize M_i, and then rescale and shift to produce a meta-analytical Z-map with mean μ and variance σ^2:

$$Z_i = (M_i - \mu)/\sqrt{(\sigma^2\, \mathbf{1}^\top \mathbf{Q} \mathbf{1}/N^2)} \times \sigma + \mu.$$

Voxelwise correction for false discovery rate (5% level) was performed using the two-stage linear step-up procedure[43].

The random-effects variance across unthresholded statistical maps of the different teams was estimated using an analog to the tau-squared statistic commonly used to assess heterogeneity in meta-analysis. We used the following estimator to account for the interstudy correlation and provide an unbiased estimate of the between-team variance,

$\tau_i^2 = Y_i' \mathbf{R} Y_i$ / tr(**RQ**),

where Y_i is the vector of T statistics across teams at a given voxel i, **Q** is the correlation matrix across teams (pooling over all voxels), and **R** is the centering matrix ($\mathbf{R} = \mathbf{I} - \mathbf{11}^T/N$); this is just the usual sample variance except N−1 is replaced by tr(**RQ**).

In the case where all results are identical, this statistic should take a value approaching zero. Median tau across teams was well above one (range across hypotheses: 1.13-1.85), and visualization of voxelwise tau maps (Extended Data Figure 3a) showed much higher variability in activated regions, with some voxels showing values greater than 5. As a point of comparison, the sampling variability of T-scores over different datasets always has a standard deviation of at least 1.0, and thus it is notable that inter-team variability on the same dataset is often substantially larger.

Prediction markets

A limitation of the prediction markets part of the study is that the number of observations for each set of prediction markets is low, as it equals the number of hypotheses (n = 9) tested by the teams with the fMRI dataset. This meant that we had nine prediction market observations for "team members" and nine prediction market observations for "non-team members". These were aggregated market observations about predictions of the fraction of teams reporting significant results for each hypothesis (bounded between 0 and 1). The low number of observations implied that the statistical power to find statistically significant effects was limited, and the test results should therefore be interpreted cautiously.

Traders self-ranked expertise. On average, participants' self-reported expertise in neuroimaging (Likert scale from 1 to 10) was 6.54 (*s.d.* = 1.93) for the "team members" prediction market and 5.98 (*s.d.* = 2.39) for the "non-team members" prediction market, respectively (Welch two-sample *t*-test: $t(173.19) = 1.77$, $p = 0.078$). The mean self-reported expertise in decision sciences (Likert scale from 1 to 10) was significantly higher for the "non-team members" (*mean* = 5.13, *s.d.* = 2.36) compared to the "team members" (*mean* = 4.23, *s.d.* = 2.46) prediction market (Welch two-sample

t-test: $t(184.97) = 2.56$, $p = 0.011$). These tests comparing the value of the variables between the two samples were not pre-registered and are included for descriptive purposes.

Exploratory analyses. Although not stated in the pre-analysis plan, we examined the correlation between participants' final payoffs, as an indicator of market performance and prediction accuracy, with participants' self-reported expertise in neuroimaging and decision sciences. The Spearman correlations between payoffs and self-rated expertise turn out to be low in magnitude and statistically insignificant for expertise in both neuroimaging ($r = 0.06$, $p = 0.45$, $n = 148$) and decision sciences ($r = -0.07$, $p = 0.369$, $n = 148$). This exploratory result also holds if we examine Spearman correlations for "team members" and "non-team members" separately (expertise in neuroimaging: "non-team members", $r = 0.19$, $p = 0.141$, $n = 65$; "team-members", $r = -0.12$, $p = 0.273$, $n = 83$; expertise in decision sciences: "non-team members", $r = 0.04$, $p = 0.745$, $n = 65$; "team-members", $r = 0.02$, $p = 0.829$, $n = 83$).

To explore whether and how market prices (i.e., market's predictions) aggregate traders' private information over time, we calculated the absolute error of the market price from the fundamental value on an hourly basis (average price of all transactions within an hour), resulting in a time series of 240 observations (10 days x 24 hours; see Extended Data Figure 5a). We ran two panel regressions with 18 cross-sections (i.e., nine hypotheses run for both sets of markets) and 240 time observations each. In the model (1), we regressed the absolute error on a binary prediction market indicator (team vs. non-team member) and control for linear time effect. The statistically significant coefficient for the team membership dummy ($\beta = -0.22$, $p < 0.001$) indicated that, on average, predictions in the "team members" prediction market were closer to the fundamental value than aggregate market's predictions in the "non-team members" prediction market. The positive coefficient for the time trend ($\beta = 4.41 \times 10^{-4}$, $p < 0.001$) in the model suggested that information aggregation got worse over time, i.e. that prices in both prediction markets tended to drift away from the fundamental value as time progressed. Adding the interaction term of the time trend and the prediction market indicator variable in model (2) revealed that prediction errors over time increased at a significantly higher rate in the "team members" prediction market compared to the "non-team members" prediction market. Despite the lower prediction errors in the "team members" prediction market, this suggests that information aggregation over time was more effective in the "non-team members" prediction market. The results are presented in Extended Data Figure 5b.

Concerning individual traders and how their opinions were incorporated in the market's predictions, we carried out two analyses for the "team members" prediction market only. First, Spearman correlations between the results their team has reported (a binary outcome) and their individual final holdings in the asset for each of the nine hypotheses range from 0.23 to 0.74 (all correlations are statistically significant, except for Hypothesis #7: $\rho_s = 0.23$, $p = 0.104$; for details, see Extended Data Table 5b). In a second analysis, we calculated the percentage of trades in the "team members" prediction markets which are consistent with the results their team reported (i.e., whether they buy when their team reported a significant result in the hypothesized direction, but the market prices reflect "no significant result" and vice versa) for each of the nine hypotheses. The fractions of consistent trades ranged from 0.68 to 0.89. One-sample Wilcoxon signed-rank tests for a share of 0.5 revealed that the share of consistent trades was significantly higher than 50% (z-values range from 2.78 to 6.81; $p < 0.004$ for all tests; see Extended Data Table 5b for details). However, it turns out that inconsistent trades are disproportionately larger (in terms of volume) than consistent trades, explaining the systematic overvaluation of fundamental values.

In order to test whether overoptimism of traders in the team prediction market was the result of over-representation of teams reporting significant results, we computed the fraction of active traders that reported a significant result for each hypothesis. Overall, active traders in the teams prediction market were representative with respect to the overall results. The absolute differences in the fraction of significant results for active traders compared to all teams are small and vary from 0.021 to 0.088. For all hypotheses, the fraction of significant results for active traders lies within the 95% confidence intervals associated with the fraction of significant results reported by all teams, indicating that the active traders' information in the market are representative for the overall results. Moreover, for all hypotheses but one (Hypothesis #5), the fraction of significant results was lower for the active traders compared to all teams. Therefore, overoptimism of the traders in the teams prediction market could not be attributed to a biased outcome for these researchers.

Supplementary Discussion

The goal and scope of NARPS

The main goal of NARPS was to ecologically test the degree of analytical variability in fMRI and the effects of this variability on analytical results. Therefore, we tried to mimic as much as possible

the analytical variability that occurs "in the wild". To this aim, we collected a real sample of a value-based decision-making fMRI task, with some complexity that would yield variance in reported significant results, similar to what is happening in standard fMRI studies in practice. Importantly, the study was not meant to determine the ultimate validity of the teams' results, which is not possible in this case due to the lack of ground truth for the hypothesized effects. The second goal of NARPS was to assess the accuracy of predictions made by researchers in the field regarding the fraction of teams reporting significant results for each hypothesis. To this aim, we employed the novel approach of prediction markets, where participants trade on the outcomes of scientific analyses[2–5].

Analytical variability and its related factors

Seventy analysis teams independently analyzed the same fMRI dataset to test the same nine ex-ante hypotheses which were based on the relevant scientific literature (Extended Data Table 1). Exploratory analyses of the relation between reported hypothesis outcomes and a subset of specific measurable analytical choices and image features identified several primary sources of analytical variability across teams. First, teams differed in the way they modelled the hypotheses (i.e. the regressors and contrasts they included in the model). Second, there were multiple different software packages used. Third, teams differed in the preprocessing steps applied as well as the parameters and techniques used at each preprocessing step. Fourth, teams differed in the threshold used to identify significant effects at each voxel in the brain and the method used to correct for multiple comparisons. Finally, teams differed in how the anatomical regions of interest (ROIs) were defined to determine whether there was a significant effect in each a priori ROI.

Reported analysis outcomes demonstrated substantial variability in results across analysis teams (Figure 1 and Extended Data Table 2). We further found that while the agreement between thresholded statistical maps was largely limited to which voxels were not activated (which comprised the large majority of voxels), correlations between the unthresholded statistical maps across teams were moderate. We performed exploratory analyses to assess the impact of specific factors on the variability of results. These analyses pointed out specific factors that significantly contributed to the variability. Higher estimated smoothness of the unthresholded statistical map, analyzing the data with FSL, and using parametric correction methods were all related to more significant results, though the latter two effects were not consistently supported by nonparametric

bootstrap analyses. While the analysis software and correction method used are analytical choices directly made by each team, the estimated smoothness is a feature of the map and is affected by multiple earlier analytical choices. For example, exploratory analysis showed that modeling head movement was related to reduced estimated smoothness. These results imply that variability in results could potentially be reduced, for instance by converging to a specific correction method or analysis software. However, we do not believe that there is an optimal software or correction method across studies and hypotheses, as each one is optimal for different purposes (cf. [44]).

Limitations

It is important to note that our analyses are limited by the number of teams and factors. Although we aimed to test the variability that is present in practice, and the cooperation of the neuroimaging community was overwhelming, our analyses have limited statistical power due to the skewed distributions of hypothesis outcomes (e.g. the low number of teams that reported insignificant results for hypotheses 7-9), and for many analytical choices there were too many different choices across teams to allow statistical modeling. For example, we did not find significant differences in results between analysis teams that chose to use the preprocessed (with fMRIPrep) shared dataset versus the teams that chose to use the raw dataset and preprocess the data by themselves. However, preprocessing includes many analytical procedures, and the effect of each specific procedure on the variability of final results was not directly tested here due to lack of power resulting from the multiple available options for each step. We would also note that our use of hierarchical models (which pool across hypotheses with varying levels of agreement) helps increase sensitivity overall. We propose that only computational simulations could have sufficient power to estimate the theoretical contribution of each analytical choice.

There are several important analytical choices that could not be directly tested here. For example, as each hypothesis was related to a specific brain region, each team was required to choose an operative definition of the specific hypothesized region (i.e., in order to decide whether a significant activation was found within this region or not). Given the exact same thresholded statistical map, different teams could potentially reach different conclusions[45]. Moreover, one of the three regions of interest in the current study was the ventromedial prefrontal cortex (vmPFC), for which there is no specific agreed-upon anatomical definition. This may have further contributed to variability across teams. However, we could not include this analytical choice in the

tested model, as there were too many distinct methods used by the teams (e.g., different atlases, Neurosynth[33], visual examination, etc.) resulting in the lack of power for testing this effect.

Another important step we could not directly measure here was the general linear model specification. For example, modelling response time (or not) could potentially affect the results; the majority of teams (44) did not do so, but there were several different methods used by the teams that did, which were different enough that they could not be collapsed into a single class for modeling. We did find several model specification errors that resulted in statistical maps that were anticorrelated with the majority of teams. While some of these errors might be related to the relative complexity of the particular task used here, other errors, such as those involving the inclusion of multiple correlated parameters in the model, likely generalize to all tasks.

The fact that correlated unthresholded statistical maps resulted in substantially different binary results across analysis teams suggested that a main source of the variability comes from the final stages of analysis: thresholding, correcting for multiple comparisons and anatomical ROI specifications. Although the general correction method used (parametric versus nonparametric) was found to be related to the final results, exploratory analysis applying a fixed threshold, correction method and anatomical ROI specification did not yield qualitatively more similar binary results compared to the reported ones (Extended Data Figure 4). Nonetheless, correlated statistical maps should not necessarily produce similar binary results when applying the same threshold, since the correlation coefficient is not sensitive to overall scaling and thus correlated values could differ substantially in magnitude. Use of consistent thresholding and meta-analytical approaches provide another view on the heterogeneity of results (Extended Data Figure 4b). Hypotheses 2, 4, 5 & 6 all had at least 50% of teams showing activation on some consistent thresholding approach as well as significant voxels in the image-based meta-analysis (IBMA).

It should also be noted that our results are conditional on the specific task we chose to use here, the mixed gambles task. This task is relatively complex, with multiple parametric modulators that could be (and were) modeled in a number of different ways. While this is a relatively representative task in the field of value-based decision-making, a simpler task may have resulted in lower variance across pipelines (e.g., if there was less flexibility in the specification of the statistical model and region definitions, and less possibility of model misspecification).

Prediction markets

We used prediction markets to test the degree to which researchers from the field can predict the results. Prediction markets are assumed to aggregate private information distributed among traders, and can generate and disseminate a consensus among market participants. While traders in the "team members" prediction market had the data and knew their own results, traders from the "non-team members" reported significantly higher expertise in decision-sciences and are therefore assumed to be more familiar with the relevant literature. Nonetheless, we found that both groups of traders strongly overestimated the fraction of significant results. These results indicate that researchers in the field are over-optimistic with regard to the reproducibility of results across analysis teams. Nonetheless, team members predicted the relative plausibility of the hypotheses very well. Surprisingly, neither self-rated expertise in neuroimaging nor self-rated expertise in decision-sciences were related to better performance in the prediction markets (i.e., to better predictions of the results; see Supplementary Methods and Results).

Implications regarding previous findings with the mixed gambles task

There is a spectrum of concerns regarding the quality of research, ranging from replicability (the ability to reproduce a result in a new sample) to computational reproducibility (the ability to reproduce a result given data and analysis plans)[46]. Concerns over replicability across many areas of science have led to a number of projects in recent years that attempted to assess the replicability of empirical findings across labs[3,34,47,48]. While such an undertaking would certainly be useful in the context of fMRI, the expense of fMRI data collection makes a large-scale replication attempt across many studies very unlikely. The present study does not broadly assess the replicability of neuroimaging research, but it does provide valuable insights, given that the design of the present study overlaps (in the equal indifference group) with the previous study of Tom et al.[10]. Out of the four primary claims made in the initial paper (reflecting significant outcomes on Hypotheses #1, #3 and #5, and a null outcome on Hypothesis #7), two were supported by a majority of teams in the present study. Moreover, as results largely differed for the equal indifference group (for which the design was similar to Tom et al.[10]) and the equal range group (for which the design was similar to De Martino et al.[11] and Canessa et al.[12]), mainly for the negative loss effect in the vmPFC (Hypothesis #5 vs. Hypothesis #6), inconsistent findings across these studies may be the result of the different designs they used. However, as the present study did not aim to directly test

replicability of fMRI findings, but rather the variability across analysis pipelines with the same dataset, the implications are limited and should be interpreted with caution.

General implications and proposed solutions

In this study, we assessed the degree to which results are reproducible across multiple analysts given a single dataset and pre-defined hypotheses. Our findings raise substantial concerns and indicate an urgent need for better understanding and controlling the effects of analytical choices on reported results. Furthermore, our findings indicate that the further one gets from raw data the more divergent the results are. One implication of these findings is that meta-analyses should be more effective when using less processed data (i.e. unthresholded statistical maps versus thresholded statistical maps or activation coordinates)[cf. 49].

Importantly, the analysis teams who participated in the present study were not incentivized to find significant effects, which is thought to drive a number of questionable research practices (e.g., "p-hacking"[7]). Furthermore, our results suggest that the teams were not consistently biased towards either affirmation or rejection of hypotheses: Several hypotheses were affirmed by roughly 5% of teams, while Hypothesis #5 was affirmed by 84% of teams. Thus, the variability in the present results more likely reflects actual variability in the standard analytical methods used by the participating research groups and their interaction with the data, as well as model specification differences and errors present for some teams. It should be noted that the analyses and results submitted by the teams were not individually peer reviewed, and we cannot know for certain whether and how the peer review process would have affected the results. In addition, in the present study all analysis teams used a univariate analysis (GLM) approach, although not explicitly instructed to do so (one team performed GLM at the first level within participant analysis, but partial least squares correlation at the group level analysis). While this type of analysis has been the most frequent one since the advent of fMRI, in recent years many studies have been using multivariate pattern analyses[50], which are less standardized and are therefore prone to be affected by specific analytical choices. An open question is how the present results would generalize to those studies in which the researchers are motivated to detect a significant result (due to the prevailing bias for publication of significant results). Our results imply substantial researcher degrees of freedom resulting in ample scope for p-hacking, as a significant result for each

hypothesis could be reported based on at least four (based on the number of teams that reported a significant result for hypotheses 7-9) of the pipelines used in practice by analysis teams.

We propose that complex datasets should be analyzed using multiple analysis pipelines, preferably by more than one research team, and the results compared to ensure concordance across validated pipelines. The current study and future ones could point at the main analytical choices that lead to variable results. "Multiverse analysis" thus can be focused on those analytical choices to save required computational resources and allow a wider use across research groups. Previous studies in other fields have suggested different versions of "multiverse" analysis[19,20], but these have yet to be widely implemented. Meta-analysis methods can be used to draw conclusions based on multiple analysis pipelines and/or studies (when unthresholded statistical maps would be shared alongside neuroimaging publications). We believe this is a promising and important future direction, given the substantial influence of analytical choices on reported results. We also propose that the use of well-engineered and well-validated software tools instead of custom solutions, when appropriate, can help reduce the presence of errors and suboptimal analysis choices simply by the fact that these have been tested by multiple users and often employ more rigorous software engineering practices (but, importantly, should not be treated as a "black box").

It is important to note, however, that concordance among different analysis pipelines does not necessarily imply that the conclusion of those analyses is correct. In the present study we chose to collect and distribute real fMRI task data of a somewhat complex value-based decision-making task, rather than synthetic data with known effects, in order to achieve the crucial ecological aspect of this project (and also to allow the use of prediction markets). Moreover, using synthetic data could have potentially introduced bias towards finding a result, as analysis teams would likely infer that some activation must be present. Since we collected a real dataset, we do not have a "ground truth" regarding the effects (i.e., we do not know for certain whether each hypothesis is correct or not). Therefore, the present study provides crucial evidence and insights regarding the variability of results across analysis pipelines in practice and its related factors, but not regarding the validity of each analytical choice or which analytical choices are the best ones. Future studies can use simulated data or null data, where the ground truth is known, to validate analysis workflows (e.g.[32,51]). These studies could potentially identify optimal analysis pipelines, on which the "multiverse analysis" could rely. We do not, however, believe that there is a single (or even a few) best analysis pipeline across studies[52,53]. Novel analysis methods are important for scientific

discovery and progress, and different pipelines are optimal for different studies and scientific questions. Therefore, we suggest to focus on "multiverse analysis", while aggregating evidence across studies by sharing unthresholded statistical maps, analysis code and design matrices, and applying meta-analysis approaches. The discussed challenges and potential solutions are relevant far beyond neuroimaging, to any scientific field where the data are complex and there are multiple acceptable analysis workflows.

Supplementary References

41. Laird, A. R. et al. ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. Hum. Brain Mapp. 25, 155–164 (2005).
42. Turkeltaub, P. E. et al. Minimizing within-experiment and within-group effects in Activation Likelihood Estimation meta-analyses. Hum. Brain Mapp. 33, 1–13 (2012).
43. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507 (2006).
44. Churchill, N. W. et al. Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. Human Brain Mapping vol. 33 609–627 (2012).
45. Hong, Y. W., Yoo, Y., Han, J., Wager, T. D. & Woo, C. W. False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. Neuroimage 195, 384–395 (2019).
46. Peng, R. D. Reproducible research in computational science. Science 334, 1226–1227 (2011).
47. Klein, R. A. et al. Investigating variation in replicability: A 'many labs' replication project. Soc. Psychol. 45, 142–152 (2014).
48. Klein, R. A. et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science 1, 443–490 (2018).
49. Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D. & Nichols, T. E. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. Neuroimage 45, 810–823 (2009).

50. Woolgar, A., Jackson, J. & Duncan, J. Coding of Visual, Auditory, Rule, and Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. J. Cogn. Neurosci. 28, 1433–1454 (2016).
51. Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G. & Rosseel, Y. neuRosim : an R package for generating fMRI data. J. Stat. Softw. 44, 1–18 (2011).
52. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18, 115–126 (2017).
53. Churchill, N. W. et al. Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity. PLoS One 7, e31147 (2012).