Article

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

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

Received: 14 November 2019

Accepted: 7 April 2020

Published online: 20 May 2020

Check for updates

A list of authors and affiliations appears in the online version of the paper.

Data analysis workflows in many scientific domains have become increasingly complex and flexible. Here we assess the effect of this flexibility on the results of functional magnetic resonance imaging by asking 70 independent teams to analyse the same dataset, testing the same 9 ex-ante hypotheses[1]. The flexibility of analytical approaches is exemplified by the fact that no two teams chose identical workflows to analyse the data. This flexibility resulted in sizeable variation in the results of hypothesis tests, even for teams whose statistical maps were highly correlated at intermediate stages of the analysis pipeline. Variation in reported results was related to several aspects of analysis methodology. Notably, a meta-analytical approach that aggregated information across teams yielded a significant consensus in activated regions. Furthermore, prediction markets of researchers in the field revealed an overestimation of the likelihood of significant findings, even by researchers with direct knowledge of the dataset[2–5]. Our findings show that analytical flexibility can have substantial effects on scientific conclusions, and identify factors that may be related to variability in the analysis of functional magnetic resonance imaging. The results emphasize the importance of validating and sharing complex analysis workflows, and demonstrate the need for performing and reporting multiple analyses of the same data. Potential approaches that could be used to mitigate issues related to analytical variability are discussed.

Data analysis workflows in many areas of science have a large number of analysis steps that involve many possible choices (that is, "researcher degrees of freedom"[6,7]). Simulation studies show that variability in analytical choices can have substantial effects on results[8], but its degree and effect in practice is unclear. Recent work in psychology addressed this through a "many analysts" approach[9], in which the same dataset was analysed by a large number of groups, uncovering substantial variability in behavioural results across analysis teams. In the Neuroimaging Analysis Replication and Prediction Study (NARPS), we applied a similar approach to the domain of functional magnetic resonance imaging (fMRI), the analysis workflows of which are complex and highly variable. Our goal was to assess—with the highest possible ecological validity—the degree and effect of analytical flexibility on fMRI results in practice. In addition, we estimated the beliefs of researchers in the field regarding the degree of variability in analysis outcomes using prediction markets to test whether peers in the field could predict the results[2–5].

Variability of results across teams

The first aim of NARPS was to assess the real-world variability of results across independent teams analysing the same dataset. The dataset included fMRI data from 108 individuals, each performing one of two versions of a task that was previously used to study decision-making under risk[10]. The two versions were designed to address a debate on the effect of gain and loss distributions on neural activity in this task[10–12]. A full description of the dataset is available in a Data Descriptor[1]; the dataset is openly available at https://doi.org/10.18112/openneuro.ds001734.v1.0.4.

Seventy teams (69 of whom had previous fMRI publications) were provided with the raw data, and an optional preprocessed version of the dataset (with fMRIPrep[13]). They were asked to analyse the data to test nine ex-ante hypotheses (Extended Data Table 1), each consisting of a description of activity in a specific brain region in relation to a particular feature of the task. They were given up to 100 days to report whether each hypothesis was supported on the basis of a whole-brain-corrected analysis (yes or no). In addition, each team submitted a detailed report of the methods of analysis that they had used, together with unthresholded and thresholded statistical maps supporting each hypothesis test (Extended Data Tables 2, 3a). To perform an ecologically valid study testing the sources of variability that contribute to published literature 'in the wild', the instructions to the teams were as minimal as possible. The only instructions were to perform the analysis as they usually would in their own research laboratory and report the binary decision on the basis of their own criteria for a whole-brain-corrected result for the specific region described in the hypothesis. The dataset, reports and collections were kept private until after the prediction markets were closed.

Overall, the rates of reported significant findings varied across hypotheses (Fig. 1, Extended Data Table 1). Only one hypothesis (hypothesis 5) showed a high rate of significant findings (84.3%), whereas three other hypotheses showed consistent non-significant findings across teams (5.7% significant findings). For the remaining five hypotheses, the results were variable, with 21.4% to 37.1% of teams reporting a significant result. The extent of the variation in results across teams was quantified by the fraction of teams that reported a result different from the majority of teams (that is, the absolute distance from consensus). On average across the 9 hypotheses, 20% of teams

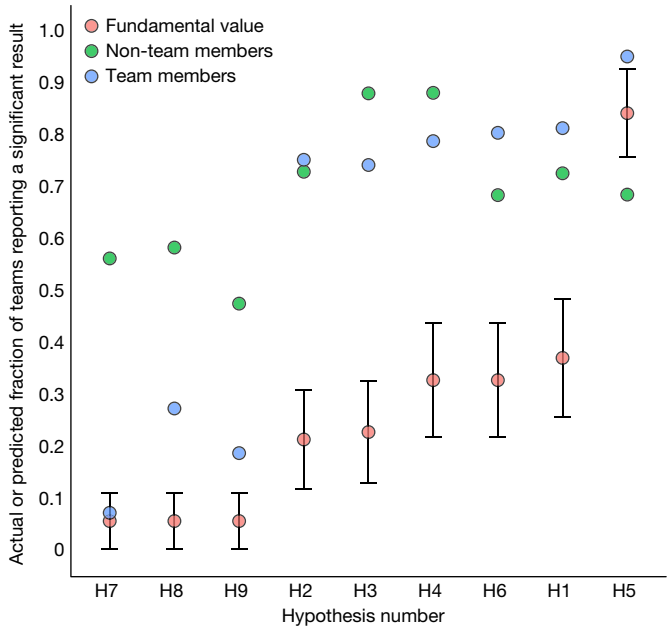


Fig. 1 | Fraction of teams reporting a significant result and prediction market beliefs. The observed fraction of teams reporting significant results (fundamental value, pink dots; $n = 70$ analysis teams), as well as final market prices for the team members markets (blue dots; $n = 83$ active traders) and the non-team members markets (green dots; $n = 65$ active traders). The corresponding 95% confidence intervals are shown for each of the nine hypotheses (note that hypotheses are sorted on the basis of the fundamental value). Confidence intervals were constructed by assuming convergence of the binomial distribution towards the normal.

reported a result that differed from the majority of teams. Given that the maximum possible variation is 50%, the observed fraction of 20% divergent results thus falls midway between complete consistency across teams and completely random results, demonstrating that analytical choices have a major effect on reported results.

Factors related to analytical variability

To examine the sources of the analytical variability in the reported binary results, we analysed the pipelines used by the teams as well as the unthresholded and thresholded statistical maps they provided. There were no two teams with identical analysis pipelines. After exclusions (Extended Data Table 3b), thresholded maps of 65 teams and unthresholded (z- or t-statistic) maps of 64 teams were included in the analyses. Fully reproducible code for all analyses of the data reported here is available at https://doi.org/10.5281/zenodo.3709273.

Variability of reported results

A set of mixed-effects logistic regression models identified several analytical variables and image features that were associated with reported outcomes (Extended Data Table 3c). The strongest factor was spatial smoothness; higher estimated smoothness of the unthresholded statistical maps (estimated using the FMRIB Software Library (FSL) smoothest function) was associated with a greater likelihood of significant outcomes ($P < 0.001$, delta pseudo-$R^2 = 0.04$; mean full width at half-maximum, 9.69 mm, range 2.50–21.28 mm across teams). Notably, although the estimated smoothness was related to the width of the applied smoothing kernel ($r = 0.71$; median applied smoothing 5 mm, range 0–9 mm across teams), the applied smoothing value itself was not significantly related to positive outcomes in a separate analysis, suggesting that the relevant smoothness arose from analytical steps beyond explicit smoothing (such as modelling of head motion; $P = 0.014$). An effect on outcomes was also found for the software package used ($P = 0.004$, delta pseudo-$R^2 = 0.04$; $n = 23$ (SPM), $n = 21$ (FSL), $n = 7$ (AFNI) and $n = 13$ (other software package))—with FSL being associated with a higher likelihood of significant results across all hypotheses compared to SPM; odds ratio = 6.69)—and for the effect of different methods of multiple test correction ($P = 0.024$, delta pseudo-$R^2 = 0.02$: $n = 48$ (parametric), $n = 14$ (nonparametric), $n = 2$ (other)), with parametric correction methods resulting in higher rates of detection than nonparametric methods. No significant effect was detected for the use of standardized preprocessed data versus custom preprocessing pipelines (48% of included teams used fMRIPrep; $P = 0.132$) or for the modelling of head motion parameters (used by 73% of the teams; $P = 0.281$). Nonparametric bootstrap analyses confirmed the significant effect of spatial smoothness, but provided inconsistent support for the effects of multiple testing and software package; because of low power, these results should be interpreted with caution.

Variability of thresholded statistical maps

The nature of analytical variability was further explored by analysing the statistical maps that were submitted by the research teams. The thresholded maps were highly sparse. Binary agreement between thresholded maps over all voxels was relatively high (median per cent agreement ranged from 93% to 99% across hypotheses), largely reflecting agreement on which voxels were not active. However, when restricted to voxels showing activation for any team, the overlap was very low (median similarity ranging from 0.00 to 0.06 across hypotheses). This may reflect variability in the number of activated voxels found by each team; for every hypothesis, the number of active voxels ranged across teams from zero to tens of thousands (Extended Data Table 4a). Analysis of the overlap between activated voxels showed that the proportion of teams with activation in the most frequently activated voxel for a given hypothesis ranged between 0.23 and 0.77 (Extended Data Fig. 1).

Variability of unthresholded statistical maps

Analysis of the correlation between unthresholded z-statistic maps across teams showed that for each hypothesis, a large cluster of teams had statistical maps that were strongly positively correlated with one another (Fig. 2, Extended Data Fig. 2). The mean Spearman correlation between all pairs of unthresholded maps (Extended Data Table 4b) was moderate (mean correlation range 0.18–0.52 across hypotheses), with higher correlations within the main cluster of analysis teams (range 0.44–0.85 across hypotheses). An analysis of voxelwise heterogeneity across unthresholded maps (equivalent to tau-squared) demonstrated that inter-team variability was large—in many cases several times the variability expected across different datasets (Extended Data Fig. 3a).

For hypotheses 1 and 3, there was a subset of seven teams whose unthresholded maps were anticorrelated with those of the main cluster of teams. A comparison of the average map for the anticorrelated cluster for hypotheses 1 and 3 confirmed that this map was highly correlated ($r = 0.87$) with the overall task-activation map, as previously reported[1]. Further analysis showed that four of these teams used models that did not properly separate the parametric effect of gain from overall task activation; because of the anticorrelation of value-system activations with task activations[14], this model mis-specification led to an anticorrelation with the parametric effects of gain. In two cases, the model included multiple regressors that were correlated with the gain parameter, which modified the interpretation of the primary gains regressor, and for one additional team, modelling details were not available.

The discrepancy between the overall correlations of unthresholded maps and the divergence of reported binary results (even within the highly correlated cluster) suggested that the variability in regional results might be due to procedures related to statistical correction for multiple comparisons and the subjective decision of teams on the

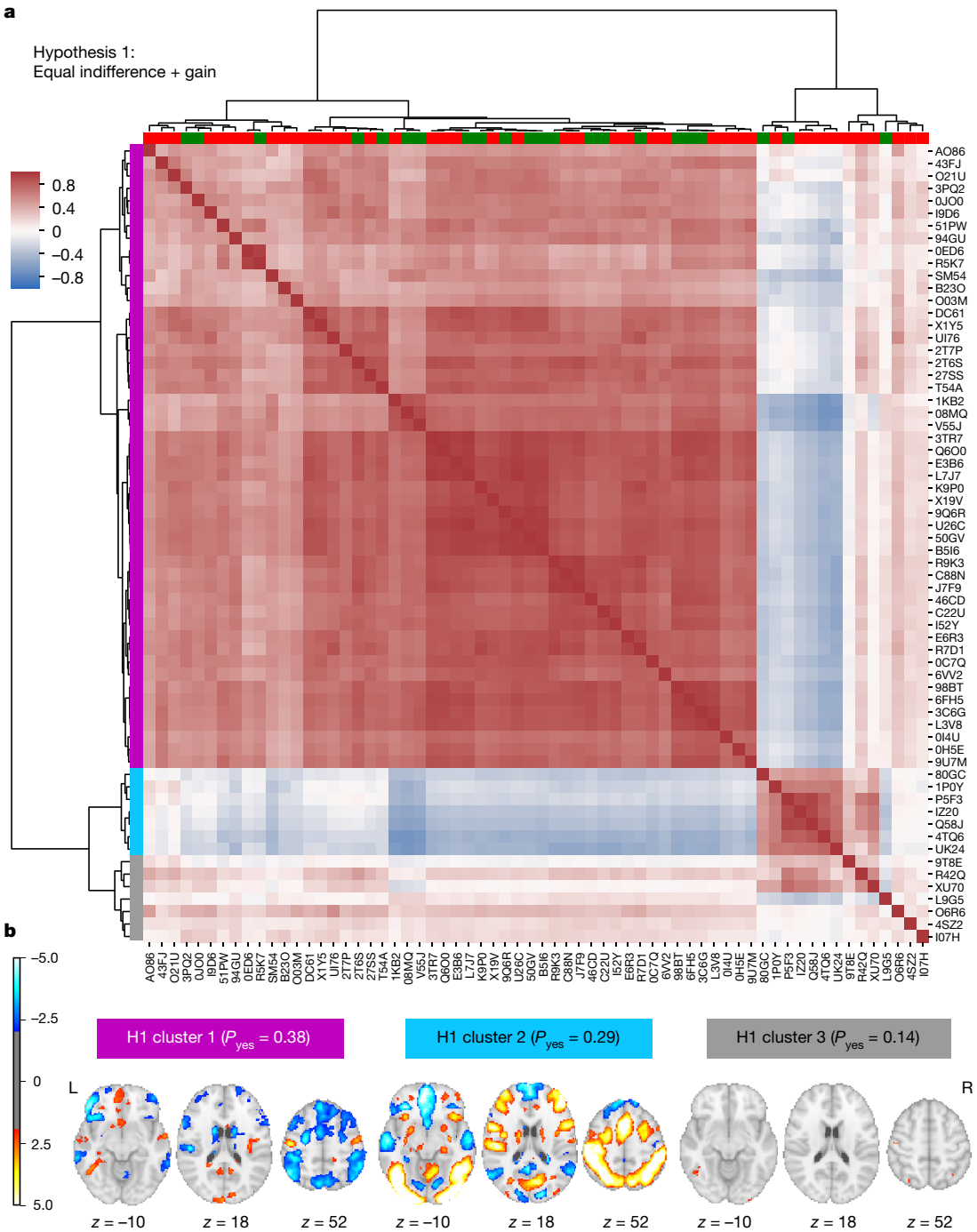


Fig. 2 | Analytical variability in whole-brain statistical results for hypothesis 1 (and hypothesis 3). a, Spearman correlation values between whole-brain unthresholded statistical maps for each team ($n = 64$) were computed and clustered according to their similarity (using Ward clustering on Euclidean distances). Row colours (left) denote cluster membership (purple, cluster 1; blue, cluster 2; grey, cluster 3); column colours (top) represent hypothesis decisions (green, yes; red, no). Brackets represent clustering. **b**, Average statistical maps (thresholded at uncorrected $z > 2.0$) for each of the three clusters shown on the left in **a**. The probability of reporting a positive hypothesis outcome (P_{yes}) is presented for each cluster. L, left; R, right. Unthresholded maps for hypotheses 1 and 3 are identical (as they both relate to the same contrast and group but different regions), and the colours represent reported results for hypothesis 1. Images can be viewed at https://identifiers.org/neurovault.collection:6048.

anatomical specification of regions of interest (ROIs). To test this, we applied a consistent thresholding method and ROI specification on the unthresholded maps across all teams for each hypothesis. This showed that even using a correction method known to be liberal and a standard anatomical definition for all regions, the degree of variability across results was qualitatively similar to that of the actual reported decisions (Extended Data Fig. 4).

We assessed the consistency across teams using an image-based meta-analysis (accounting for correlations due to common data), which demonstrated significant active voxels for all hypotheses except for hypothesis 9 after false discovery rate (FDR) correction (Extended Data Fig. 3b) and confirmatory evidence for hypotheses 2, 4, 5 and 6. These results show that inconsistent results at the individual team level underlie consistent results when the results of teams are combined.

Prediction markets

The second aim of NARPS was to test whether peers in the field could predict the results, using prediction markets in which researchers trade on the outcomes of scientific analyses and receive monetary payouts based on performance. Prediction markets have been used to assess the replicability of scientific hypotheses in the social sciences, and have revealed correlations between market prices and actual scientific outcomes[2–5]. We performed two separate prediction markets: one involving members from analysis teams ('team members' market) and another independent market for researchers who had not participated in the analysis ('non-team members' market). The markets were open for 10 consecutive days approximately 1.5 months after all analysis teams had submitted their results (which were kept confidential). On each market, traders were provided with tokens worth US$50, and traded via an online market platform on the fraction of teams that reported a significant result for each hypothesis (that is, the fundamental values). The market prices serve as measures of the aggregate beliefs of traders for the fraction of teams reporting a significant result for each hypothesis. Overall, $n = 65$ traders actively traded in the non-team members market and $n = 83$ traded in the team members market. After the markets closed, traders were paid on the basis of their performance in the markets. The analysis of the markets was preregistered on the Open Science Framework (OSF) (https://osf.io/59ksz/). Note that because some analyses were performed on the final market prices (that is, the predictions of the markets), for which there is one value per hypothesis per market, the number of observations for each of the markets was low ($n = 9$), leading to limited statistical power. Therefore, the results should be interpreted with caution.

The predictions of the markets ranged from 0.073 to 0.952 ($m = 0.599$, s.d. = 0.325) in the team members market and from 0.476 to 0.882 ($m = 0.690$, s.d. = 0.137) in the non-team members market. Except for hypothesis 7 in the team members market, all predictions were outside the 95% confidence intervals of the fundamental values (Fig. 1, Extended Data Table 5a). The Spearman correlation between the fundamental values and the predictions of the markets was significant for the team members market ($r = 0.962$, $P < 0.001$, $n = 9$) but not for the non-team members market ($r = 0.553$, $P = 0.122$, $n = 9$), nor between the predictions of both markets ($r = 0.500$, $P = 0.170$, $n = 9$).

Wilcoxon signed-rank tests suggested that traders in both markets systematically overestimated the fundamental values (team members: $z = 2.886$, $P = 0.004$, $n = 9$; non-team members: $z = 2.660$, $P = 0.008$, $n = 9$). The result in the team members market was not driven by an overrepresentation of teams who reported significant results (Supplementary Methods and Supplementary Results). Predictions in the team members market did not significantly differ from those in the non-team members market (Wilcoxon signed-rank test, $z = 1.035$, $P = 0.301$, $n = 9$), but as mentioned above, statistical power for this test was limited. Team members generally traded in the direction consistent with the results of their own team (Extended Data Table 5b), which may explain why their collective predictions were more accurate than those of non-team members (Fig. 1). Additional results are presented in the Supplementary Information (see also Extended Data Fig. 5, Extended Data Table 5).

Discussion

The analysis of a single fMRI dataset by 70 independent analysis teams, all of whom used different analysis pipelines, revealed substantial variability in reported binary results, with high levels of disagreement across teams for most of the tested hypotheses. For every hypothesis, at least four different analysis pipelines could be found that were used in practice by research groups in the field and resulted in a significant outcome. Our findings highlight the fact that it is hard to estimate the reproducibility of single studies that are performed using a single analysis pipeline. Notably, analyses of the underlying statistical parametric maps on which the hypothesis tests were based revealed greater consistency than would be expected from those inferences, and significant consensus in activated regions across teams was observed using meta-analysis. Teams with highly correlated underlying unthresholded statistical maps nonetheless reported different hypothesis outcomes (Fig. 2). Detailed analysis of the workflow descriptions and statistical results that were submitted by the analysis teams identified several common analytical variables that were related to differential reporting of significant outcomes, including the spatial smoothness of the data (a result of multiple factors beyond the applied smoothing kernel), the choice of analysis software and the correction method; however, the last two were not consistently supported by nonparametric bootstrap analyses. In addition, we identified model-specification errors for several analysis teams, which led to statistical maps that were anticorrelated with the majority for some of the hypotheses. Prediction markets that were performed on the outcomes of analyses demonstrated a general overestimation by researchers of the likelihood of significant results across hypotheses—even by those researchers who had analysed the data themselves—reflecting a marked optimism bias by researchers in the field.

The substantial amount of analytical variability, and the subsequent variability of reported hypothesis results with the same data, demonstrates that steps need to be taken to improve the reproducibility of data analysis outcomes. First, we suggest that unthresholded statistical maps should be shared as a standard practice alongside thresholded statistical maps using tools such as NeuroVault[15]. In the long run, the shared maps will allow the use of image-based meta-analysis, which we found to provide converging results across laboratories. Second, public sharing of data and analysis code should become common practice, to enable others to run their own analysis with the same data or to validate the code used. These practices, combined with the use of preregistration[16] or registered reports[17], will reduce researcher degrees of freedom but would not prevent analytical variability, as demonstrated here; however, they would ensure that the effects of variability can be assessed. All of the data and code used in the current study are publicly available with a fully reproducible execution environment for all figures and results. We believe that this can serve as an example for future studies.

Foremost, we propose that complex datasets should be analysed using several analysis pipelines, and preferably by more than one research team. Achieving such 'multiverse analysis' on a large scale will require the development of automated statistical analysis tools (for example, FitLins[18]) that can run a broad range of pipelines and assess their convergence. Different versions of such multiverse analysis have been suggested in other fields[19–21], but are not widely used. Analysis pipelines should also be validated using simulated data to assess their validity with regard to ground truth, and assessed for their effects on predictions with new data[22].

Our findings emphasize the urgent need to develop new practices and tools to overcome the challenge of variability across analysis pipelines and its effect on analytical results. Nonetheless, we maintain that fMRI can provide reliable answers to scientific questions, as strongly demonstrated in the meta-analytical results across teams along with numerous large-scale studies in the literature and replication of many findings using fMRI. Moreover, although the present investigation was limited to the analysis of a single fMRI dataset, it seems highly likely that similar variability will be present for other fields of research in which the data are high dimensional and the analysis workflows are complex and varied. The multiverse approach combined with meta-analysis is suggested as a promising solution. Notably, transparent scientific projects that involve community-wide self-assessment—such as this one—are definitive evidence of the awareness of researchers of reproducibility concerns, and the desire to assess their effect and improve practices accordingly (for additional discussion see Supplementary Discussion).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2314-9.

1. Botvinik-Nezer, R. et al. fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study. *Sci. Data* **6**, 106 (2019).
2. Dreber, A. et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl Acad. Sci. USA* **112**, 15343–15347 (2015).
3. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
4. Camerer, C. F. et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
5. Forsell, E. et al. Predicting replication outcomes in the Many Labs 2 study. *J. Econ. Psychol.* **75**, 102117 (2019).
6. Wicherts, J. M. et al. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid P-hacking. *Front. Psychol.* **7**, 1832 (2016).
7. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
8. Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6**, 149 (2012).
9. Silberzahn, R. et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. Adv. *Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
10. Tom, S. M., Fox, C. R., Trepel, C. & Poldrack, R. A. The neural basis of loss aversion in decision-making under risk. *Science* **315**, 515–518 (2007).
11. De Martino, B., Camerer, C. F. & Adolphs, R. Amygdala damage eliminates monetary loss aversion. *Proc. Natl Acad. Sci. USA* **107**, 3788–3792 (2010).
12. Canessa, N. et al. The functional and structural neural basis of individual differences in loss aversion. *J. Neurosci.* **33**, 14307–14317 (2013).
13. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
14. Acikalin, M. Y., Gorgolewski, K. J. & Poldrack, R. A. A coordinate-based meta-analysis of overlaps in regional specialization and functional connectivity across subjective value and default mode networks. *Front. Neurosci.* **11**, 1 (2017).
15. Gorgolewski, K. J. et al. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* **9**, 8 (2015).
16. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
17. Nosek, B. A. & Lakens, D. Registered reports: a method to increase the credibility of published results. *Soc. Psychol.* **45**, 137–141 (2014).
18. Markiewicz, C., De La Vega, A., Yarkoni, T., Poldrack, R. & Gorgolewski, K. FitLins: reproducible model estimation for fMRI. Poster W621 in *25th Annual Meeting of the Organization for Human Brain Mapping* (OHBM, 2019).
19. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve: descriptive and inferential statistics on all reasonable specifications. https://doi.org/10.2139/ssrn.2694998 (2015).
20. Patel, C. J., Burford, B. & Ioannidis, J. P. A. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* **68**, 1046–1058 (2015).
21. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
22. LaConte, S. et al. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. *Neuroimage* **18**, 10–27 (2003).

Rotem Botvinik-Nezer[1,2,3], Felix Holzmeister[4], Colin F. Camerer[5], Anna Dreber[6,7], Juergen Huber[4], Magnus Johannesson[6], Michael Kirchler[4], Roni Iwanir[1,2], Jeanette A. Mumford[8], R. Alison Adcock[9,10], Paolo Avesani[11,12], Blazej M. Baczkowski[13], Aahana Bajracharya[14], Leah Bakst[15,16], Sheryl Ball[17,18], Marco Barilari[19], Nadège Bault[20], Derek Beaton[21], Julia Beitner[22,23], Roland G. Benoit[24], Ruud M. W. J. Berkers[24], Jamil P. Bhanji[25], Bharat B. Biswal[26,27], Sebastian Bobadilla-Suarez[28], Tiago Bortolini[29], Katherine L. Bottenhorn[30], Alexander Bowring[31], Senne Braem[32,33], Hayley R. Brooks[34], Emily G. Brudner[25], Cristian B. Calderon[32], Julia A. Camilleri[35,36], Jaime J. Castrellon[9,37], Luca Cecchetti[38], Edna C. Cieslik[35,36], Zachary J. Cole[39], Olivier Collignon[12,19], Robert W. Cox[40], William A. Cunningham[41], Stefan Czoschke[42], Kamalaker Dadi[43], Charles P. Davis[44,45,46], Alberto De Luca[47], Mauricio R. Delgado[25], Lysia Demetriou[48,49], Jeffrey B. Dennison[50], Xin Di[26,27], Erin W. Dickie[51,52], Ekaterina Dobryakova[53], Claire L. Donnat[54], Juergen Dukart[35,36], Niall W. Duncan[55,56], Joke Durnez[57], Amr Eed[58], Simon B. Eickhoff[35,36], Andrew Erhart[34], Laura Fontanesi[59], G. Matthew Fricke[60], Shiguang Fu[61,62], Adriana Galván[63], Remi Gau[19], Sarah Genon[35,36], Tristan Glatard[64], Enrico Glerean[65], Jelle J. Goeman[66], Sergej A. E. Golowin[55], Carlos González-García[32], Krzysztof J. Gorgolewski[67], Cheryl L. Grady[21], Mikella A. Green[9,37], João F. Guassi Moreira[63], Olivia Guest[28,68], Shabnam Hakimi[9], J. Paul Hamilton[69], Roeland Hancock[45,46], Giacomo Handjaras[38], Bronson B. Harry[70], Colin Hawco[71], Peer Herholz[72], Gabrielle Herman[71], Stephan Heunis[73,74], Felix Hoffstaedter[35,36], Jeremy Hogeveen[75,76], Susan Holmes[54], Chuan-Peng Hu[77], Scott A. Huettel[37], Matthew E. Hughes[78], Vittorio Iacovella[12], Alexandru D. Iordan[79], Peder M. Isager[80], Ayse I. Isik[81], Andrew Jahn[82], Matthew R. Johnson[39,83], Tom Johnstone[78], Michael J. E. Joseph[71], Anthony C. Juliano[84], Joseph W. Kable[85,86], Michalis Kassinopoulos[87], Cemal Koba[38], Xiang-Zhen Kong[88], Timothy R. Koscik[89], Nuri Erkut Kucukboyaci[53,90], Brice A. Kuhl[91], Sebastian Kupek[92], Angela R. Laird[93], Claus Lamm[94,95], Robert Langner[35,36], Nina Lauharatanahirun[96,97], Hongmi Lee[98], Sangil Lee[85], Alexander Leemans[47], Andrea Leo[38], Elise Lesage[32], Flora Li[99,100], Monica Y. C. Li[44,45,46,101], Phui Cheng Lim[39,83], Evan N. Lintz[39], Schuyler W. Liphardt[102], Annabel B. Losecaat Vermeer[94], Bradley C. Love[28,103], Michael L. Mack[41], Norberto Malpica[104], Theo Marins[29], Camille Maumet[105], Kelsey McDonald[37], Joseph T. McGuire[15,16], Helena Melero[104,106,107], Adriana S. Méndez Leal[63], Benjamin Meyer[77,108], Kristin N. Meyer[109], Glad Mihai[110,111], Georgios D. Mitsis[112], Jorge Moll[29,67], Dylan M. Nielson[113], Gustav Nilsonne[114,115], Michael P. Notter[116], Emanuele Olivetti[11,12], Adrian I. Onicas[38], Paolo Papale[38,117], Kaustubh R. Patil[35,36], Jonathan E. Peelle[14], Alexandre Pérez[72], Doris Pischedda[118,119,120], Jean-Baptiste Poline[72,121], Yanina Prystauka[44,45,46], Shruti Ray[26], Patricia A. Reuter-Lorenz[79], Richard C. Reynolds[122], Emiliano Ricciardi[38], Jenny R. Rieck[21], Anais M. Rodriguez-Thompson[109], Anthony Romyn[41], Taylor Salo[30], Gregory R. Samanez-Larkin[9,37], Emilio Sanz-Morales[104], Margaret L. Schlichting[41], Douglas H. Schultz[39,83], Qiang Shen[61,62], Margaret A. Sheridan[109], Jennifer A. Silvers[63], Kenny Skagerlund[123,124], Alec Smith[17,18], David V. Smith[50], Peter Sokol-Hessner[34], Simon R. Steinkamp[125], Sarah M. Tashjian[63], Bertrand Thirion[43], John N. Thorp[126], Gustav Tinghög[127,128], Loreen Tisdall[67,129], Steven H. Tompson[96], Claudio Toro-Serey[15,16], Juan Jesus Torre Tresols[43], Leonardo Tozzi[130], Vuong Truong[55,56], Luca Turella[12], Anna E. van 't Veer[131], Tom Verguts[32], Jean M. Vettel[132,133,134], Sagana Vijayarajah[41], Khoi Vo[9,37], Matthew B. Wall[135,136,137], Wouter D. Weeda[131], Susanne Weis[35,36], David J. White[138], David Wisniewski[32], Alba Xifra-Porxas[87], Emily A. Yearling[44,45,46], Sangsuk Yoon[139], Rui Yuan[130], Kenneth S. L. Yuen[77,108], Lei Zhang[94], Xu Zhang[45,46,140], Joshua E. Zosky[39,83], Thomas E. Nichols[31] ✉, Russell A. Poldrack[67] ✉ & Tom Schonberg[1,2] ✉

[1]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel. [2]Department of Neurobiology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [4]Department of Banking and Finance, University of Innsbruck, Innsbruck, Austria. [5]HSS and CNS, California Institute of Technology, Pasadena, CA, USA. [6]Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [7]Department of Economics, University of Innsbruck, Innsbruck, Austria. [8]Center for Healthy Minds, University of Wisconsin–Madison, Madison, WI, USA. [9]Center for Cognitive Neuroscience, Duke University, Durham, NC, USA. [10]Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC, USA. [11]Neuroinformatics Laboratory, Fondazione Bruno Kessler, Trento, Italy. [12]Center for Mind/Brain Sciences - CIMeC, University of Trento, Rovereto, Italy. [13]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. [14]Department of Otolaryngology, Washington University in St. Louis, St. Louis, MO, USA. [15]Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA. [16]Center for Systems Neuroscience, Boston University, Boston, MA, USA. [17]Department of Economics, Virginia Tech, Blacksburg, VA, USA. [18]School of Neuroscience, Virginia Tech, Blacksburg, VA, USA. [19]Crossmodal Perception and Plasticity Laboratory, Institutes for Research in Psychology (IPSY) and Neurosciences (IoNS), UCLouvain, Louvain-la-Neuve, Belgium. [20]School of Psychology, University of Plymouth, Plymouth, UK. [21]Rotman Research Institute, Baycrest Health Sciences Centre, Toronto, Ontario, Canada. [22]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands. [23]Department of Psychology, Goethe University, Frankfurt am Main, Germany. [24]Max Planck Research Group: Adaptive Memory, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. [25]Department of Psychology, Rutgers University–Newark, Newark, NJ, USA. [26]Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, USA. [27]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. [28]Department of Experimental Psychology, University College London, London, UK. [29]D'Or Institute for Research and Education (IDOR), Rio de Janeiro, Brazil. [30]Department of Psychology, Florida International University, Miami, FL, USA. [31]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [32]Department of Experimental Psychology, Ghent University, Ghent, Belgium. [33]Department of Psychology, Vrije Universiteit Brussel, Brussels, Belgium. [34]Department of Psychology, University of Denver, Denver, CO, USA. [35]Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, Jülich, Germany. [36]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [37]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA. [38]MoMiLab Research Unit, IMT School for Advanced Studies Lucca, Lucca, Italy. [39]Department of Psychology, University of Nebraska–Lincoln, Lincoln, NE, USA. [40]National Institute of Mental Health (NIMH), National Institutes of Health, Bethesda, MD, USA. [41]Department of Psychology, University of Toronto, Toronto, Ontario, Canada. [42]Institute of Medical Psychology, Goethe University, Frankfurt am Main, Germany. [43]Inria, CEA, Université Paris-Saclay, Palaiseau, France. [44]Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA. [45]Brain Imaging Research Center, University of Connecticut, Storrs, CT, USA. [46]Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT, USA. [47]PROVIDI Lab, Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands. [48]Section of Endocrinology and Investigative Medicine, Faculty of Medicine, Imperial College London, London, UK. [49]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK. [50]Department of Psychology, Temple University, Philadelphia, PA, USA. [51]Krembil Centre for Neuroinformatics, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [52]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [53]Center for Traumatic Brain Injury Research, Kessler Foundation, East Hanover, NJ, USA. [54]Department of Statistics, Stanford University, Stanford, CA, USA. [55]Graduate Institute of Mind, Brain and Consciousness, Taipei Medical University, Taipei, Taiwan. [56]Brain and Consciousness Research Centre, TMU-ShuangHo Hospital, New Taipei City, Taiwan. [57]Department of Psychology and Stanford Center for Reproducible Neuroscience, Stanford University, Stanford, CA, USA. [58]Instituto de Neurociencias, CSIC-UMH, Alicante, Spain. [59]Faculty of Psychology, University of Basel, Basel, Switzerland. [60]Computer Science Department, University of New Mexico, Albuquerque, NM, USA. [61]School of Management, Zhejiang University of Technology, Hangzhou, China. [62]Institute of Neuromanagement, Zhejiang University of Technology, Hangzhou, China. [63]Department of Psychology, University of California Los Angeles, Los Angeles, CA, USA. [64]Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada. [65]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland. [66]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [67]Department of Psychology, Stanford University, Stanford, CA, USA. [68]Research Centre on Interactive Media, Smart Systems and Emerging Technologies - RISE, Nicosia, Cyprus. [69]Center for Social and Affective Neuroscience, Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden. [70]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, New South Wales, Australia. [71]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [72]McConnell Brain Imaging Centre, The Neuro (Montreal Neurological Institute-Hospital), Faculty of Medicine, McGill University, Montreal, Quebec, Canada. [73]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. [74]Department of Research and Development, Epilepsy Centre Kempenhaeghe, Heeze, The Netherlands. [75]Department of Psychology, University of New Mexico, Albuquerque, NM, USA. [76]Psychology Clinical Neuroscience Center, University of New Mexico, Albuquerque, NM, USA. [77]Leibniz-Institut für Resilienzforschung (LIR), Mainz, Germany. [78]School of Health Sciences, Swinburne University of Technology, Hawthorn, Victoria, Australia. [79]Department of Psychology, University of Michigan, Ann Arbor, MI, USA. [80]Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands. [81]Department of Neuroscience, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany. [82]fMRI Laboratory, University of Michigan, Ann Arbor, MI, USA. [83]Center for Brain, Biology and Behavior, University of Nebraska–Lincoln, Lincoln, NE, USA. [84]Center for Neuropsychology and Neuroscience Research, Kessler Foundation, East Hanover, NJ, USA. [85]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA. [86]MindCORE, University of Pennsylvania, Philadelphia, PA, USA. [87]Graduate Program in Biological and Biomedical Engineering, McGill University, Montreal, Quebec, Canada. [88]Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [89]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, IA, USA. [90]Department of Physical Medicine and Rehabilitation, Rutgers New Jersey Medical School, Newark, NJ, USA. [91]Department of Psychology, University of Oregon, Eugene, OR, USA. [92]Faculty of Economics and Statistics, University of Innsbruck, Innsbruck, Austria. [93]Department of Physics, Florida International University, Miami, Florida, USA. [94]Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria. [95]Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria. [96]US CCDC Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, MD, USA. [97]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA. [98]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA. [99]Fralin Biomedical Research Institute, Roanoke, VA, USA. [100]Economics Experimental Lab, Nanjing Audit University, Nanjing, China. [101]Haskins Laboratories, New Haven, CT, USA. [102]Biology Department, University of New Mexico, Albuquerque, NM, USA. [103]The Alan Turing Institute, London, UK. [104]Laboratorio de Análisis de Imagen Médica y Biometría (LAIMBIO), Universidad Rey Juan Carlos, Madrid, Spain. [105]Inria, Univ Rennes, CNRS, Inserm, IRISA UMR 6074, Empenn ERL U 1228, Rennes, France. [106]Departamento de Psicobiología, División de

Psicología, CES Cardenal Cisneros, Madrid, Spain. [107]Northeastern University Biomedical Imaging Center, Northeastern University, Boston, MA, USA. [108]Neuroimaging Center (NIC), Focus Program Translational Neurosciences (FTN), Johannes Gutenberg University Medical Center Mainz, Mainz, Germany. [109]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [110]Max Planck Research Group: Neural Mechanisms of Human Communication, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. [111]Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, Dresden, Germany. [112]Department of Bioengineering, McGill University, Montreal, Quebec, Canada. [113]Data Science and Sharing Team, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA. [114]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden. [115]Department of Psychology, Stockholm University, Stockholm, Sweden. [116]The Laboratory for Investigative Neurophysiology (The LINE), Department of Radiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland. [117]Department of Vision and Cognition, Netherlands Institute for Neuroscience, Amsterdam, The Netherlands. [118]Bernstein Center for Computational Neuroscience and Berlin Center for Advanced Neuroimaging and Clinic for Neurology, Charité Universitätsmedizin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany. [119]Cluster of Excellence Science of Intelligence, Technische Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany. [120]NeuroMI - Milan Center for Neuroscience, Milan, Italy. [121]Henry H. Wheeler, Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, CA, USA. [122]Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA. [123]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden. [124]Center for Social and Affective Neuroscience, Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden. [125]Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Research Centre Jülich, Jülich, Germany. [126]Department of Psychology, Columbia University, New York, NY, USA. [127]Department of Management and Engineering, Linköping University, Linköping, Sweden. [128]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden. [129]Center for Cognitive and Decision Sciences, University of Basel, Basel, Switzerland. [130]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA. [131]Methodology and Statistics Unit, Institute of Psychology, Leiden University, Leiden, The Netherlands. [132]US Combat Capabilities Development Command Army Research Laboratory, Aberdeen, MD, USA. [133]University of California Santa Barbara, Santa Barbara, CA, USA. [134]University of Pennsylvania, Philadelphia, PA, USA. [135]Invicro, London, UK. [136]Faculty of Medicine, Imperial College London, London, UK. [137]Clinical Psychopharmacology Unit, University College London, London, UK. [138]Centre for Human Psychopharmacology, Swinburne University, Hawthorn, Victoria, Australia. [139]Department of Management and Marketing, School of Business, University of Dayton, Dayton, OH, USA. [140]Biomedical Engineering Department, University of Connecticut, Storrs, CT, USA. ✉e-mail: thomas.nichols@bdi.ox.ac.uk; poldrack@stanford.edu; schonberg@tauex.tau.ac.il

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

fMRI dataset

To test the variability of neuroimaging results across analysis pipelines used in practice in research laboratories, we distributed a single fMRI dataset to independent analysis groups from around the world, requesting them to test nine predefined hypotheses. The full dataset is publicly available in the Brain Imaging Data Structure (BIDS)[23] on OpenNeuro (https://doi.org/10.18112/openneuro.ds001734.v1.0.4) and is described in detail in a Data Descriptor[1].

In brief, the fMRI dataset consisted of data from 108 participants who performed a mixed gambles task, which is often used to study decision-making under risk. In this task, participants are asked on each trial to accept or reject a presented prospect. The prospects consist of an equal 50% chance of either gaining a given amount of money or losing another, similar or different, amount of money. Participants were divided into two groups: in the 'equal indifference' group ($n = 54$) the potential losses were half the size of the potential gains[10] (reflecting the 'loss aversion' phenomenon, in which people tend to be more sensitive to losses than to equal-sized gains[24]); and in the 'equal range' group ($n = 54$) the potential losses and the potential gains were taken from the same scale[11,12]. The two groups were used to resolve inconsistencies of previous published results.

The dataset was distributed to the teams via Globus (https://www.globus.org/). The distributed dataset included raw data of 108 participants ($n = 54$ for each experimental group), as well as the same data after preprocessing with fMRIPrep v.1.1.4 (RRID: SCR_016216)[13]. The fMRIPrep preprocessing mainly included brain extraction, spatial normalization, surface reconstruction, head motion estimation and susceptibility distortion correction. Both the raw and the preprocessed datasets underwent quality assurance (described in detail in the Data Descriptor[1]).

MRI data collection was approved by the Helsinki committee at Sheba Tel Hashomer Medical Center and the ethics committee at Tel Aviv University, and all participants gave written informed consent (as described in the Data Descriptor of this dataset[1]). The Board for Ethical Questions in Science at the University of Innsbruck approved the data collection in the prediction markets, and certified that the project complied with all requirements of the ethical principles and guidelines of good scientific practice. The Stanford University Institutional Review Board (IRB) determined that the analysis of the submitted team results did not meet the definition of human subject research, and thus no further IRB review was required. We have complied with all relevant ethical regulations.

Predefined hypotheses

Previous studies with the mixed gambles task suggested that activity in the ventromedial prefrontal cortex and ventral striatum, among other brain regions, is related to the magnitude of the potential gain[10]. A fundamental open question in the field of decision-making under risk is whether the magnitude of the potential loss is coded by the same brain regions (through negative activation), or by regions related to negative emotions, such as the amygdala[10–12]. The specific hypotheses included in NARPS were chosen to address this open question, using two different designs that were used in those previous studies (that is, equal indifference versus equal range). Each analysis team tested the same nine predefined hypotheses (Extended Data Table 1). Each hypothesis predicted fMRI activations in a specific brain region, in relation to a specific aspect of the task (gain or loss amount) and a specific group (equal indifference or equal range, or a comparison between the two groups). Therefore, for each hypothesis, the maximum sample size was 54 participants (hypotheses 1–8) or 54 participants per group in the group comparison (hypothesis 9). Although the hypotheses referred to specific brain regions, analysis teams were instructed to report their results on the basis of a whole-brain analysis (not an ROI-based analysis, as is sometimes used in fMRI studies).

Recruitment of and instructions to analysis teams

We recruited analysis teams via social media, mainly Twitter and Facebook, as well as during the 2018 annual meeting of the Society for Neuroeconomics. Ninety-seven teams registered to participate in the study. Each team consisted of up to three members. To ensure independent analyses across teams, and to prevent influencing the subsequent prediction markets, all team members signed an electronic nondisclosure agreement that they would not release, publicize or discuss their results with anyone until the end of the study. All team members of 82 teams signed the nondisclosure form. They were offered co-authorship on the present publication in return for their participation.

Analysis teams were provided with access to the full dataset. They were asked to freely analyse the data with their usual analysis pipeline to test the nine hypotheses and report a binary decision for each hypothesis on whether it was significantly supported on the basis of a whole-brain analysis. Although the hypotheses were region-specific, we clearly requested a whole-brain analysis result to avoid the need of teams to create and share masks of ROIs. Each team also filled in a full report of the analysis methods used (following the guidelines of the Committee on Best Practices in Data Analysis and Sharing; COBIDAS[25]) and created a collection on NeuroVault[15] (RRID: SCR_003806) with one unthresholded and one thresholded statistical map for each hypothesis, on which their decisions were based (teams could optionally include additional maps in their collection; see Extended Data Table 3a for links for collections). For each result (that is, the binary decision on whether a given hypothesis was supported by the data or not), teams further reported how confident they were in this result and how similar they thought their result was to the results of the other teams (each measure was an integer between 1 (not at all) to 10 (extremely)). These measures are presented in Extended Data Tables 1, 2. To measure the variability of results in an ecological manner, instructions to the analysis teams were minimized and the teams were asked to perform the analysis as they usually would in their own laboratory and to report the binary decision on the basis of their own criteria.

Seventy of the 82 teams submitted their results and reports by the final deadline (15 March 2019; overall teams were given up to 100 days, varying based on the date they joined, to complete and report their analysis). The dataset, reports and collections were kept private until the end of the study and closure of the prediction markets. To avoid identification of the teams, each team was provided with a unique random four-character ID.

Overall, 180 participants were part of NARPS analysis teams. Out of 70 analysis teams, 5 teams consisted of 1 member, 20 teams consisted of 2 members and 45 teams consisted of 3 members. Out of the 180 team members, there were 62 principal investigators, 43 post-doctoral researchers, 53 graduate students and 22 members from other positions (for example, data scientists or research analysts).

Factors related to analytical variability

To explore the factors related to the variability in results across teams, the reports of all teams were manually annotated to create a table describing the methods used by each team. Code for all analyses of the reports and statistical maps submitted by the analysis teams is openly shared in GitHub (https://github.com/poldrack/narps). Analyses reported in this manuscript were performed using code release v.2.0.3 (https://doi.org/10.5281/zenodo.3709273). We performed exploratory analyses of the relation between the reported hypothesis outcomes and several analytical choices and image features using mixed-effects

logistic regression models implemented in R, with the lme4 package[26]. The factors included in the model were: hypothesis number, estimated smoothness (based on the smoothest function in FSL), use of standardized preprocessing, software package, method of correction for multiple comparisons and modelling of head movement. The teams were modelled as a random effect. One team submitted results that were not based on a whole-brain analysis as requested, and therefore their data were excluded from all analyses.

Inferences using logistic regression models were confirmed using nonparametric bootstrap analysis, resampling data team-wise to maintain random effect structure. For the continuous or binary regressors (smoothness, movement modelling and use of fMRIPrep data), we computed bootstrap confidence intervals and, as an approximate hypothesis test, tested whether the confidence interval includes zero. For the factorial variables (hypothesis, software package and multiple testing method), this was not possible because there is not a single coefficient for the factor; in addition, for software package and multiple testing methods, some bootstrap samples did not contain all values of the factor. For these variables we instead performed model comparison between the full model and a reduced model excluding each factor, and computed the proportion of times the full model was selected on the basis of the model selection criterion (using both Bayesian information criterion and Akaike information criterion) being numerically lower in the full model[27].

In addition, we performed exploratory analyses to examine the variability across statistical maps submitted by the teams. The unthresholded and thresholded statistical maps of all teams were resampled to common space (FSL MNI space, 91 × 109 × 91, 2 mm isotropic) using nilearn[28] (RRID: SCR_001362). For unthresholded maps, we used third-order spline interpolation; for thresholded maps, we used linear interpolation and then thresholded at 0.5, to prevent artefacts that appeared when using nearest neighbour interpolation. Of the 69 teams included in the analyses, unthresholded maps of 5 teams and thresholded maps of 4 teams were excluded from the image-based analyses (see Extended Data Table 3b for details). As some of the hypotheses reflected negative activations—which can be represented by either positive or negative values in the statistical maps, depending on the model used—we asked the teams to report the direction of the values in their maps for the relevant hypotheses (5, 6 and 9). Unthresholded maps were corrected to address sign flips for reversed contrasts as reported by the analysis teams. In addition, *t* values were converted to *z* values with Hughett's transform[29]. All subsequent analyses of the unthresholded maps were performed only on voxels that contained non-zero data for all teams (range across hypotheses: 111,062–145,521 voxels).

We assessed the agreement between thresholded statistical maps using per cent agreement, that is, the per cent of voxels that have the same binary value. Because the thresholded maps are very sparse, these values are necessarily high when computed across all voxels. Therefore, we also computed the agreement between pairs of statistical maps only for voxels that were non-zero for at least one member of each pair. To further test the agreement across teams, we performed a coordinate-based meta-analysis with activation likelihood estimation[30,31] (see Supplementary Information).

We further computed the correlation between the unthresholded images of the 64 teams. The correlation matrices were clustered using Ward clustering; the number of clusters was set to three for all hypotheses on the basis of visual examination of the dendrograms. A separate mean statistical map was then created for the teams in each cluster (see Fig. 2, Extended Data Fig. 2). Drivers of map similarity were further assessed by modelling the median correlation distance of each team from the average pattern as a function of several analysis decisions (for example, smoothing, whether or not the data preprocessed with fMRIPrep were used, and so on).

To assess the effect of variability in thresholding methods and anatomical definitions across teams, unthresholded *z* maps for each team were thresholded using a common approach. The *z* maps for each team were translated to *P* values, which were then thresholded using two approaches: a heuristic correction (known to be liberal[32]), and a voxelwise FDR correction. Note that it was not possible to compute the commonly used familywise error correction using Gaussian random field theory because residual smoothness was not available for each team. We then identified whether there were any suprathreshold voxels within the appropriate anatomical ROI for each hypothesis. The ROIs for the ventral striatum and amygdala were defined anatomically on the basis of the Harvard-Oxford anatomical atlas. As there is no anatomical definition for the ventromedial prefrontal cortex, we defined the region using a conjunction of anatomical regions (including all anatomical regions in the Harvard-Oxford atlas that overlap with the ventromedial portion of the prefrontal cortex) and a meta-analytical map obtained from https://neurosynth.org/ (ref. [33]) for the search term "ventromedial prefrontal".

An image-based meta-analysis was used to quantify the evidence for each hypothesis across analysis teams (Extended Data Fig. 3b), accounting for the lack of independence due to the use of a common dataset across teams. See Supplementary Information for a description of the image-based meta-analysis method.

Prediction markets

The second main goal of NARPS was to test the degree to which researchers in the field can predict results, using prediction markets[2–5,34]. We invited team members (researchers that were members of one of the analysis teams) and non-team members (researchers that were neither members of any of the analysis teams nor members of the NARPS research group) to participate in a prediction market[2,35] to measure peer beliefs about the fraction of teams reporting significant whole-brain-corrected results for each of the nine hypotheses. The prediction markets were conducted 1.5 months after all teams had submitted their analysis of the fMRI dataset. Thus, team members had information about the results of their specific team, but not about the results of any other team.

Similar to previous studies[2–5], participants in the prediction markets were provided with monetary endowments (100 tokens, worth US$50) and traded on the outcome of the hypotheses through a dedicated online market platform. Each hypothesis constitutes one asset in the market, with asset prices predicting the fraction of teams reporting significant whole-brain-corrected results for the corresponding ex-ante hypothesis examined by the analysis teams using the same dataset. Trading on the prediction markets was incentivized, that is, traders were paid on the basis of their performance in the markets.

Recruitment. For the non-team members prediction market, we invited participants via social media (mainly Facebook and Twitter) and emails. The invitation contained a link to an online form on the NARPS website (https://www.narps.info/) where participants could sign up using their email address.

Participants for the team members prediction market were invited, after all teams submitted their results, by an email that directed them to an independent registration form (with identical form fields), to separate participants for the two prediction markets already at the time of registration. Note that team members were not aware to start with that they would be invited to participate in a separate prediction market after they had analysed the data. The decision to implement a second market, consisting of traders with partial information about the fundamental values (that is, the team members) was made after the teams obtained access to the fMRI dataset. Thus, team members were only invited to participate in the market after all teams had submitted their analysis results. Once the registration for participating in the prediction markets had been closed, we reconciled the sign-ups with the list of team members to ensure that team members did not mistakenly end up in the non-team members prediction market and vice versa.

In addition to their email addresses, which were used as the only key to match registrations, accounts in the market platform and the teams' analysis results, registrants were required to provide the following information during sign-up: (i) name, (ii) affiliation, (iii) position (PhD candidate, post-doctoral researcher, assistant professor, senior lecturer, associate professor, full professor, other), (iv) years since PhD, (v) gender, (vi) age, (vii) country of residence, (viii) self-assessed expertise in neuroimaging (Likert scale ranging from 1 to 10), (ix) self-assessed expertise in decision sciences (Likert scale ranging from 1 to 10), (x) preferred mode of payment (Amazon.de voucher, Amazon.com voucher, PayPal payment), and (xi) whether they are a team member of any analysis team (yes or no). The invitations to participate in the prediction markets were first distributed on 9 April 2019; the registration closed on 29 April 2019 at 16:00 UTC. Once registration closed, all participants received a personalized email containing a link to the web-based market software and their login credentials. The prediction markets opened on 2 May 2019 at 16:00 UTC and closed on 12 May 2019 at 16:00 UTC.

Information available to participants. All participants had access to detailed information about the data collection, the experimental protocol, the ex-ante hypotheses, the instructions given to the analysis teams, references to related papers and detailed instructions about the prediction markets via the NARPS website (https://www.narps.info/).

Implementation of prediction markets. To implement the prediction markets, we used a newly developed web-based framework dedicated for conducting continuous-time online market experiments, inspired by the trading platform in the Experimental Economics Replication Project (EERP)[3] and the Social Sciences Replication Project (SSRP)[4]. Similar to these previous implementations, there were two main views on the platform: (i) the market overview and (ii) the trading interface. The market overview showed the nine assets (that is, one corresponding to each hypothesis) in tabular format, including information on the (approximate) current price for buying a share and the number of shares held (separated for long and short positions) for each of the nine hypotheses. Via the trading interface, which was shown after clicking on any of the hypotheses, the participant could make investment decisions and view price developments for the particular asset.

Note that initially, there was an error in the labelling of two assets (that is, hypotheses) in the trading interface and the overview table of the web-based trading platform (the more detailed hypothesis description available via the info symbol on the right-hand side of the overview table contained the correct information): hypotheses 7 and 8 mistakenly referred to negative rather than positive effects of losses in the amygdala. One of the participants informed us about the inconsistency between the information on the trading interface and the information provided on the website on 6 May 2019. The error was corrected immediately on the same day and all participants were informed about the mistake on our part through a personal email notification (on 6 May 2019, 15:28 UTC), pointing out explicitly which information was affected and asking them to double-check their holdings in the two assets to make sure that they were invested in the intended direction.

Trading and market pricing. In both prediction markets, traders were endowed with 100 tokens (the experimental currency unit). Once the markets opened, these tokens could be used to trade shares in the assets (that is, hypotheses). Unlike prediction markets on binary outcomes (for example, the outcomes of replications as in previous studies[3,4]), for which market prices were typically interpreted as the predicted probability of the outcome to occur[36] (although see two previous studies for caveats[37,38]), the prediction markets accompanying the team analyses in the current study were implemented in terms of vote-share-markets. Hence, the prediction market prices serve as measures of the aggregate beliefs of traders for the fraction of teams reporting that the hypotheses were supported and can fluctuate between 0 (no team reported a significant result) and 1 (all teams reported a significant result).

Prices were determined by an automated market maker implementing a logarithmic market scoring rule[39]. At the beginning of the markets, all assets were valued at a price of 0.50 tokens per share. The market maker calculated the price of a share for each infinitesimal transaction and updated the price on the basis of the scoring rule. This ensured both that trades were always possible even when there was no other participant with whom to trade and that participants had incentives to invest according to their beliefs[40]. The logarithmic scoring rule uses the net sales (shares held − shares borrowed) that the market maker has done so far in a market to determine the price for an infinitesimal trade as $p=e^{s/b}/(e^{s/b}+1)$. The parameter b determines the liquidity provided by the market maker and controls how strongly the market price is affected by a trade. We set the liquidity parameter to $b=100$, implying that by investing 10 tokens, traders could move the price of a single asset from 0.50 to about 0.55.

Investment decisions for a particular hypothesis were made from the market's trading interface. In the trading overview, participants could see the (approximate) price of a new share, the number of shares they currently held (separated for long and short positions) and the number of tokens their current position was worth if they liquidated their shares. The trading page also contained a graph depicting previous price developments. To make an adjustment to their current position, participants could choose either to increase or decrease their position by a number of tokens of their choice. The trading procedures and market pricing are described in more detail in a previous study[3].

Incentivization. Once the markets had been closed, the true 'fundamental value' for each asset (that is, the fraction of teams that reported a significant result for the particular hypothesis) was determined and gains and losses were calculated as follows: if holdings in a particular asset were positive (that is, the trader acted as a net buyer), the payout was calculated as the fraction of analysis teams reporting a significant result for the associated hypothesis multiplied by the number of shares held in the particular asset; if a trader's holdings were negative (that is, the trader acted as a net seller), the (absolute) amount of shares held was valued at the price differential between 1 and the fraction of teams reporting a significant result for the associated hypothesis.

Any tokens that had not been invested into shares when the market closed were voided. Any tokens awarded as a result of holding shares were converted to US dollars at a rate of 1 token = US$0.5. The final payments were transferred to participants during the months May to September 2019 in form of Amazon.com gift cards, Amazon.de gift cards or PayPal payments, depending on the preferred mode of payment indicated by the participants after registration for the prediction markets.

Participants. In total, 96 team members and 91 non-team members signed up to participate in the prediction markets. $n=83$ team members and $n=65$ non-team members actively participated in the markets. The number of traders active in each of the assets (that is, hypotheses) ranged from 46 to 76 ($m=56.4$, s.d. = 8.9) in the team members set of markets and from 35 to 58 ($m=47.1$, s.d. = 7.9) in the non-team members set of markets. See Extended Data Table 5c for data about trading volume on the prediction markets.

Of the participants, 10.2% did not work in academia (but hold a PhD), 34.2% were PhD students, 43.3% were post-doctoral researchers or assistant professors, 7.5% were lecturers or associate professors and 4.8% were full professors. 27.8% of the participants were female. The average time spent in academia after obtaining the PhD was 4.1 years. Most of the participants lived in Europe (46.3%) and North America (46.3%).

Preregistration. All analyses of the prediction markets data reported were preregistered at https://osf.io/pqeb6/. The preregistration was

completed after the markets opened, but before the markets closed. Only one member of the NARPS research group, F. Holzmeister, had any information about the prediction market prices before the markets closed (as he monitored the prediction markets). He was not involved in writing the preregistration. Only two members of the NARPS research group, R.B.-N. and T. Schonberg, had any information about the results reported by the 70 analysis teams before the prediction markets closed. Neither of them were involved in writing the preregistration. For additional details on the prediction markets, see Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The full fMRI dataset is publicly available on OpenNeuro (https://doi.org/10.18112/openneuro.ds001734.v1.0.4) and is described in detail in a Data Descriptor[1]. The results reported by all teams are presented in Extended Data Table 2. A table describing the methods used by the analysis teams is available with the analysis code. NeuroVault collections containing the submitted statistical maps are available via the links provided in Extended Data Table 3a. Source data for Figs. 1, 2 are provided with the paper. Readers may obtain access to the data and run the full analysis stream on the team submissions by following the directions at https://github.com/poldrack/narps/tree/master/ImageAnalyses. Access to the raw data requires specifying a URL for the dataset, which is: https://zenodo.org/record/3528329/files/narps_origdata_1.0.tgz. Results (automatically generated figures, results and output logs) for image analyses are available for anonymous download at https://doi.org/10.5281/zenodo.3709275.

Code availability

Code for all analyses of the reports and statistical maps submitted by the analysis teams is openly shared in GitHub (https://github.com/poldrack/narps). Image-analysis code was implemented within a Docker container, with software versions pinned for reproducible execution (https://hub.docker.com/r/poldrack/narps-analysis/tags). Python code was automatically tested for quality using the flake8 static analysis tool and the codacy.com code quality assessment tool, and the results of the image-analysis workflow were validated using simulated data. The image-analysis code was independently reviewed by an expert who was not involved in writing the original code. Prediction market analyses were performed using R v.3.6.1; packages were installed using the checkpoint package, which reproducibly installs all package versions as of a specified date (13 August 2019). Analyses reported in this manuscript were performed using code release v.2.0.3 (https://doi.org/10.5281/zenodo.3709273). Although not required to, several analysis teams publicly shared their analysis code. Extended Data Table 3d includes these teams along with the link to their code.

23. Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3**, 160044 (2016).
24. Tversky, A. & Kahneman, D. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
25. Nichols, T. E. et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* **20**, 299–303 (2017).
26. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
27. Lubke, G. H. et al. Assessing model selection uncertainty using a bootstrap approach: an update. *Struct. Equ. Modeling* **24**, 230–245 (2017).
28. Abraham, A. et al. Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* **8**, 14 (2014).
29. Hughett, P. Accurate computation of the *F*-to-*z* and *t*-to-*z* transforms for large arguments. *J. Stat. Softw.* **23**, 1–5 (2007).
30. Turkeltaub, P. E., Eden, G. F., Jones, K. M. & Zeffiro, T. A. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* **16**, 765–780 (2002).
31. Eickhoff, S. B. et al. Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage* **137**, 70–85 (2016).
32. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl Acad. Sci. USA* **113**, 7900–7905 (2016).
33. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
34. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
35. Arrow, K. J. et al. Economics. The promise of prediction markets. *Science* **320**, 877–878 (2008).
36. Wolfers, J. & Zitzewitz, E. Interpreting prediction market prices as probabilities. https://doi.org/10.3386/w12200 (NBER, 2006).
37. Manski, C. F. Interpreting the predictions of prediction markets. *Econ. Lett.* **91**, 425–429 (2006).
38. Fountain, J. & Harrison, G. W. What do prediction markets predict? *Appl. Econ. Lett.* **18**, 267–272 (2011).
39. Hanson, R. Logarithmic market scoring rules for modular combinatorial information aggregation. *J. Prediction Markets* **1**, 3–15 (2007).
40. Chen, Y. *Markets as an Information Aggregation Mechanism for Decision Support.* PhD thesis, Penn State Univ. (2005).

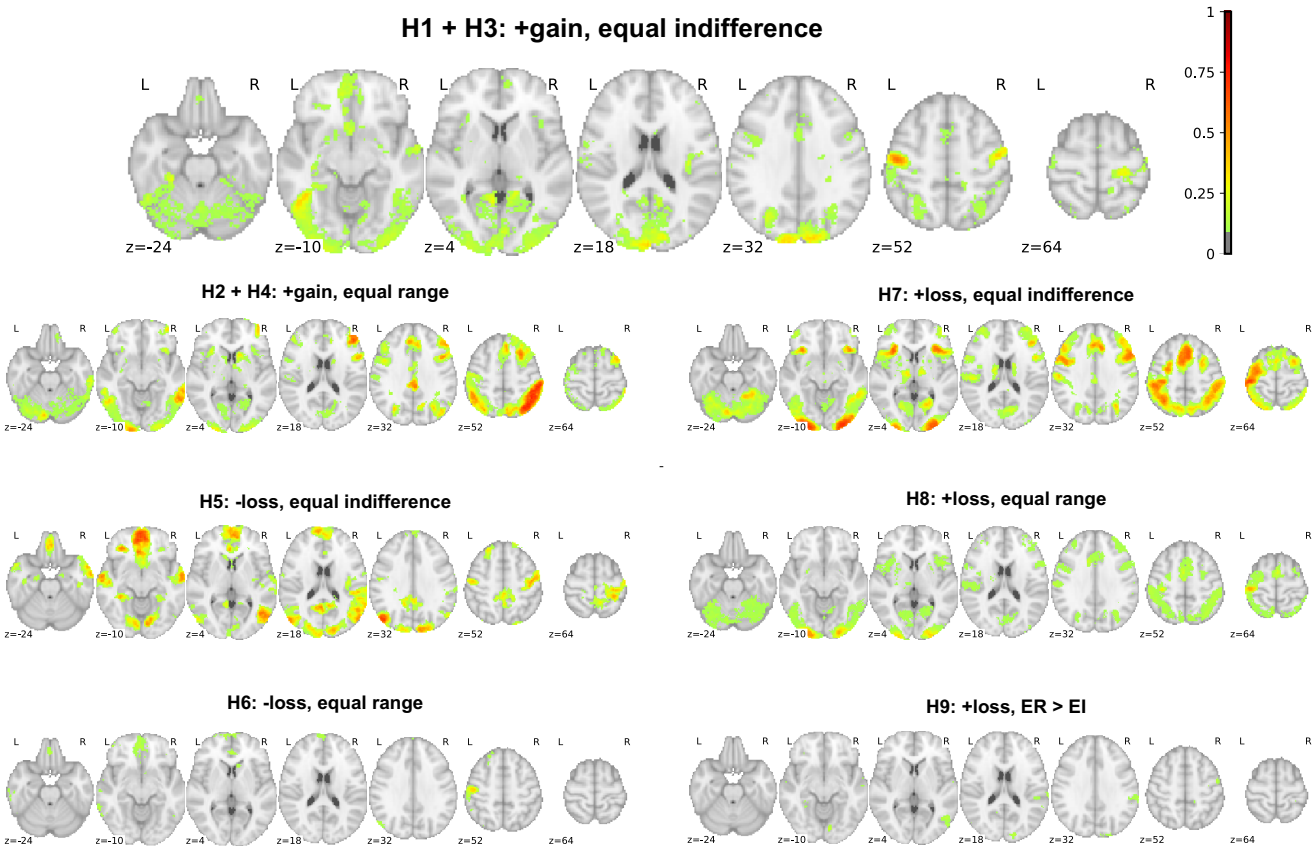
Acknowledgements Neuroimaging data collection, performed at Tel Aviv University, was supported by the Austrian Science Fund (P29362-G27), the Israel Science Foundation (ISF 2004/15 to T. Schonberg) and the Swedish Foundation for Humanities and Social Sciences (NHS14-1719:1). Hosting of the data on OpenNeuro was supported by a National Institutes of Health (NIH) grant (R24MH117179). We thank M. C. Frank, Y. Assaf and N. Daw for comments on an earlier draft; the Texas Advanced Computing Center for providing computing resources for preprocessing of the data; the Stanford Research Computing Facility for hosting the data; and D. Roll for assisting with data processing. T. Schonberg thanks The Alfredo Federico Strauss Center for Computational Neuroimaging at Tel Aviv University; A.D. thanks the Knut and Alice Wallenberg Foundation and the Marianne and Marcus Wallenberg Foundation (A.D. is a Wallenberg Scholar), the Austrian Science Fund (FWF, SFB F63) and the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser); F. Holzmeister, J. Huber and M. Kirchler thank the Austrian Science Fund (FWF, SFB F63); D.W. was supported by the Research Foundation Flanders (FWO) and the European Union's Horizon 2020 research and innovation programme (https://ec.europa.eu/programmes/horizon2020/en) under the Marie Skłodowska-Curie grant agreement no. 665501; L. Tisdall was supported by the University of Basel Research Fund for Junior Researchers; C.B.C. was supported by grant 12O7719N from the Research Foundation Flanders; E.L. was supported by grant 12T2517N from the Research Foundation Flanders and Marie Skłodowska-Curie Actions under COFUND grant agreement 665501; A. Eed was supported by a predoctoral fellowship La Caixa-Severo Ochoa from Obra Social La Caixa and also acknowledges Comunidad de Cálculo Científico del CSIC for the high-performance computing (HPC) use; C.L. was supported by the Vienna Science and Technology Fund (WWTF VRG13-007) and Austrian Science Fund (FWF P 32686); A.B.L.V. was supported by the Vienna Science and Technology Fund (WWTF VRG13-007); L.Z. was supported by the Vienna Science and Technology Fund (WWTF VRG13-007), the National Natural Science Foundation of China (no. 71801110), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (no. 18YJC630268) and China Postdoctoral Science Foundation (no. 2018M633270); D.P. is currently supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy 'Science of Intelligence' (EXC 2002/1, project number 390523135); P.H. was supported in part by funding provided by Brain Canada, in partnership with Health Canada, for the Canadian Open Neuroscience Platform initiative; J.-B.P. was partially funded by the NIH (NIH-NIBIB P41 EB019936 (ReproNim), NIH-NIMH R01 MH083320 (CANDIShare) and NIH RF1 MH120021 (NIDM)) and the National Institute Of Mental Health of the NIH under award number R01MH096906 (Neurosynth), as well as the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative and the Brain Canada Foundation with support from Health Canada; S.B.E. was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 785907 (HBP SGA2); G.M. was supported by the Max Planck Society; S. Heunis has received funding from the Dutch foundation LSH-TKI (grant LSHM16053-SGF); J.F.G.M. was supported by a Graduate Research Fellowship from the NSF and T32 Predoctoral Fellowship from the NIH; B.M. was supported by the Deutsche Forschungsgemeinschaft (grant CRC1193, subproject B01); A.R.L. was supported by NSF 1631325 and NIH R01 DA041353; M.E.H., T.J. and D.J.W. were supported by the Australian National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability; P.M.I. was supported by VIDI grant 452-17-013 from the Netherlands Organisation for Scientific Research; B.M.B. was supported by the Max Planck Society; J.P.H. was supported by a grant from the Swedish Research Council; R.W.C. and R.C.R. were supported by NIH IRP project number ZICMH002888; D.M.N., R.W.C., and R.C.R. used the computational resources of the National Institutes of Health High Performance Computing Biowulf cluster (http://hpc.nih.gov); D.M.N. was supported by NIH IRP project number ZICMH002960; C.F.C. was supported by the Tianqiao and Chrissy Center for Social and Decision Neuroscience Center Leadership Chair; R.G.B. was supported by the Max Planck Society; R.M.W.J.B. was supported by the Max Planck Society; M.B., O.C. and R.G. were supported by the Belgian Excellence of Science program (EOS project 30991544) from the FNRS-Belgium; O.C. is a research associate at the FRS-FNRS of Belgium; A.D.L. was supported by grant R4195 "Repimpact" of EraNET Neuron; Q.S. was funded by grant no. 71971199,71602175 and 71942004 from the National Natural Science Foundation of China and no. 16YJC630103 of the Ministry of Education of Humanities and Social Science; and T.E.N. was supported by the Wellcome Trust award 100309/Z/12/Z.

Author contributions NARPS management team: R.B.-N., F. Holzmeister, C.F.C., A.D., J. Huber, M.J., M. Kirchler, R.A.P. and T. Schonberg. fMRI dataset (experiment design): R.I., J. Durnez, R.A.P. and T. Schonberg. fMRI dataset (data collection): R.I. and T. Schonberg. fMRI dataset (preprocessing, quality assurance and data sharing): R.B.-N., K.J.G., R.A.P. and T. Schonberg.

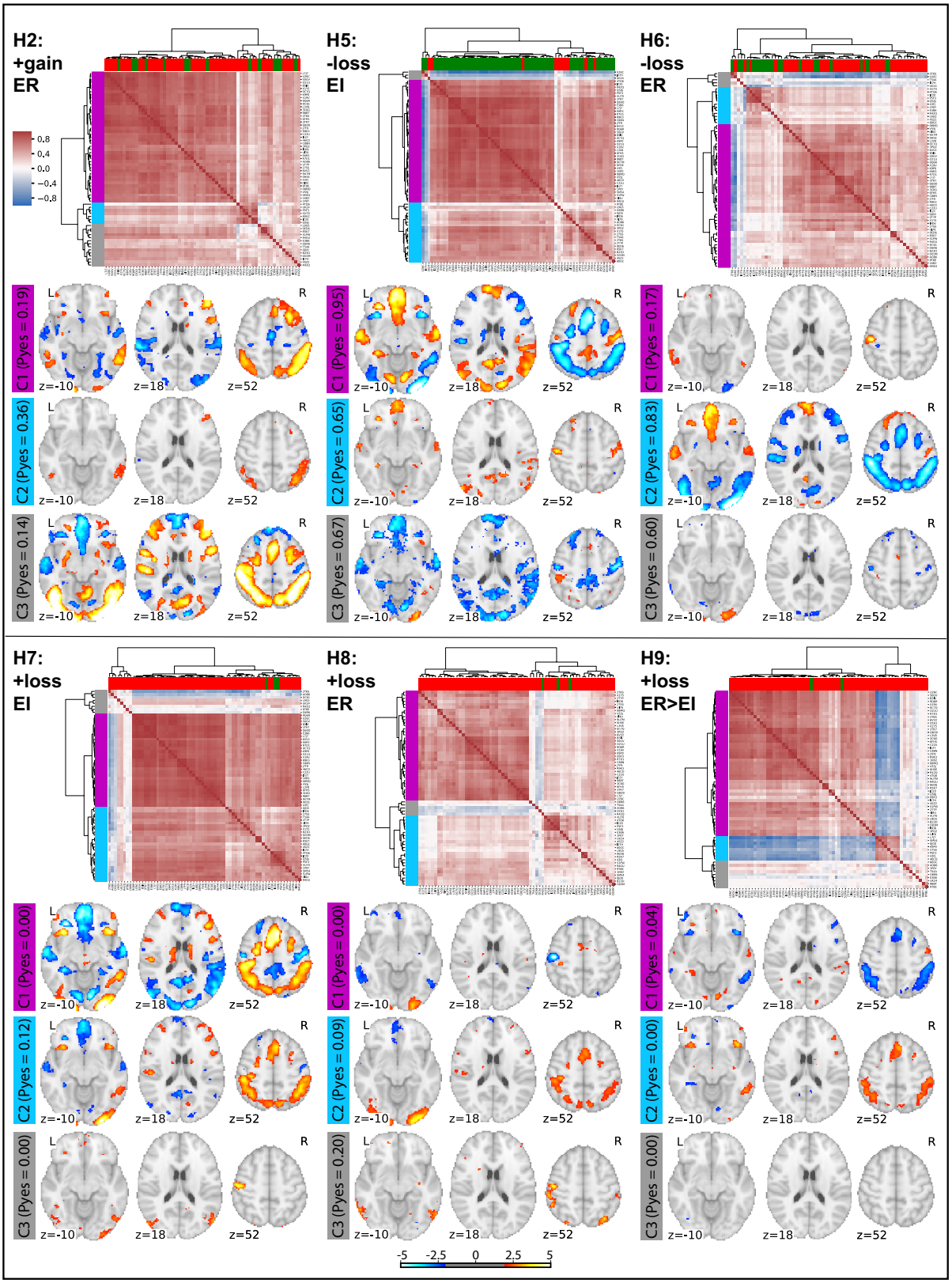
Analysis teams (recruitment, point of contact and management): R.B.-N., R.A.P. and T. Schonberg. Analysis teams (analysis of the submitted results and statistical maps): R.A.P., T.E.N., J.A.M., J.-B.P., A.P., R.B.-N. and T. Schonberg. Code review: T.G. and K.D. Prediction markets (design and management): F. Holzmeister, C.F.C., A.D., J. Huber, M.J. and M. Kirchler. Prediction markets (analysis): F. Holzmeister, R.B.-N., C.F.C., A.D., J. Huber, M.J., M. Kirchler, S.K., R.A.P. and T. Schonberg. Writing the manuscript: R.B.-N., F. Holzmeister, A.D., J. Huber, M.J., M. Kirchler, T.E.N., R.A.P. and T. Schonberg. Participated as members of analysis teams and reviewed and edited the manuscript: R.A.A., P.A., B.M.B., A. Bajracharya, L.B., S. Ball, M.B., N.B., D.B., J.B., R.G.B., R.M.W.J.B., J.P.B., B.B.B., S.B.-S., T.B., K.L.B., A. Bowring, S. Braem, H.R.B., E.G.B., C.B.C., J.A.C., J.J.C., L.C., E.C.C., Z.J.C., O.C., R.W.C., W.A.C., S.C., K.D., C.P.D., A.D.L., M.R.D., L.D., J.B.D., X.D., E.W.D., E.D., C.L.D., J. Dukart, N.W.D., A. Eed, S.B.E., A. Erhart, L.F., G.M.F., S.F., A.G., R.G., S.G., E.G., J.J.G., S.A.E.G., C.G.-G., K.J.G., C.L.G., M.A.G., J.F.G.M., O.G., S. Hakimi, J.P.H., R.H., G. Handjaras, B.B.H., C.H., P.H., G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P.H., S.A.H., M.E.H., V.I., A.D.I., P.M.I., A.I.I., A.J., M.R.J., T.J., M.J.E.J., A.C.J., J.W.K., M. Kassinopoulos, C.K., X.-Z.K., T.R.K., N.E.K., B.A.K., A.R.L., C.L., R.L., N.L., H.L., S.L., A. Leemans, A. Leo, E.L., F.L., M.Y.C.L., P.C.L., E.N.L., S.W.L., A.B.L.V., B.C.L., M.L.M., N.M., T.M., C.M., K.M., J.T.M., H.M., A.S.M.L., B.M., K.N.M., G.M., G.D.M., J.M., T.E.N., D.M.N., G.N., M.P.N., E.O., A.I.O., P.P., K.R.P., J.E.P., D.P., Y.P., S.R., P.A.R.-L., R.C.R., E.R., J.R.R., A.M.R.-T., A.R., T. Salo, G.R.S.-L., E.S.-M., M.L.S., D.H.S., Q.S., M.A.S., J.A.S., K.S., A.S., D.V.S., P.S.-H., S.R.S., S.M.T., B.T., J.N.T., G.T., L. Tisdall, S.H.T., C.T.-S., J.J.T.T., L. Tozzi, V.T., L. Turella, A.E.v.V., T.V., J.M.V., S.V., K.V., M.B.W., W.D.W., S.W., D.J.W., D.W., A.X.-P., E.A.Y., S.Y., R.Y., K.S.L.Y., L.Z., X.Z. and J.E.Z.

Competing interests The authors declare no competing interests.

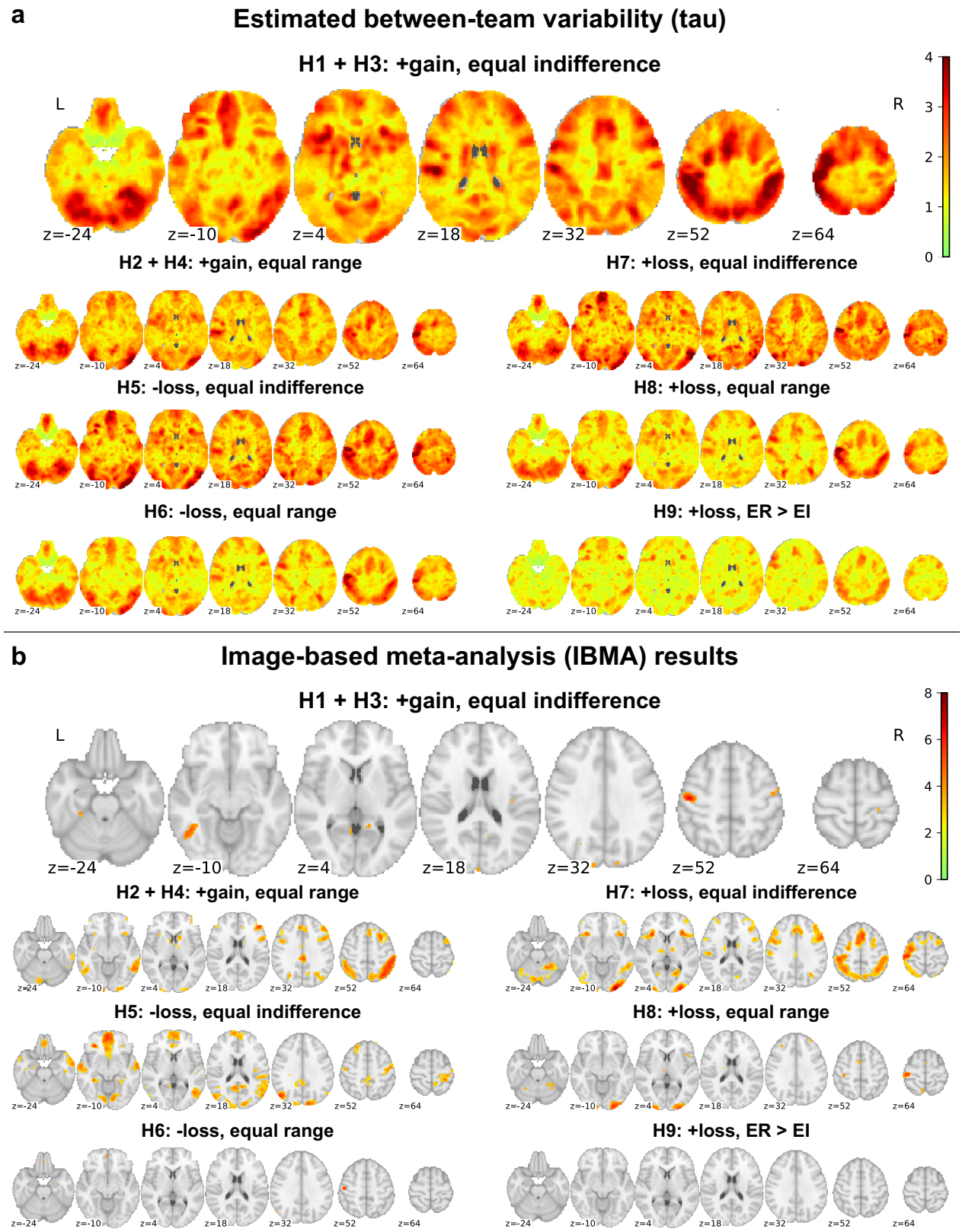
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2314-9.
Correspondence and requests for materials should be addressed to T.E.N., R.A.P. or T.S.
Peer review information *Nature* thanks Martin Lindquist, Marcus Munafo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Voxels overlap. Maps showing at each voxel the proportion of teams (out of $n = 65$ teams) that reported significant activations in their thresholded statistical map, for each hypothesis (labelled H1–H9), thresholded at 10% (that is, voxels with no colour were significant in fewer than 10% of teams). + or − refers to the direction of effect; gain or loss refers to the effect being tested; and equal indifference (EI) or equal range (ER) refers to the group being examined or compared. Hypotheses 1 and 3, as well as hypotheses 2 and 4, share the same statistical maps as they relate to the same contrast and experimental group but different regions (see Extended Data Table 1). Images can be viewed at https://identifiers.org/neurovault.collection:6047.

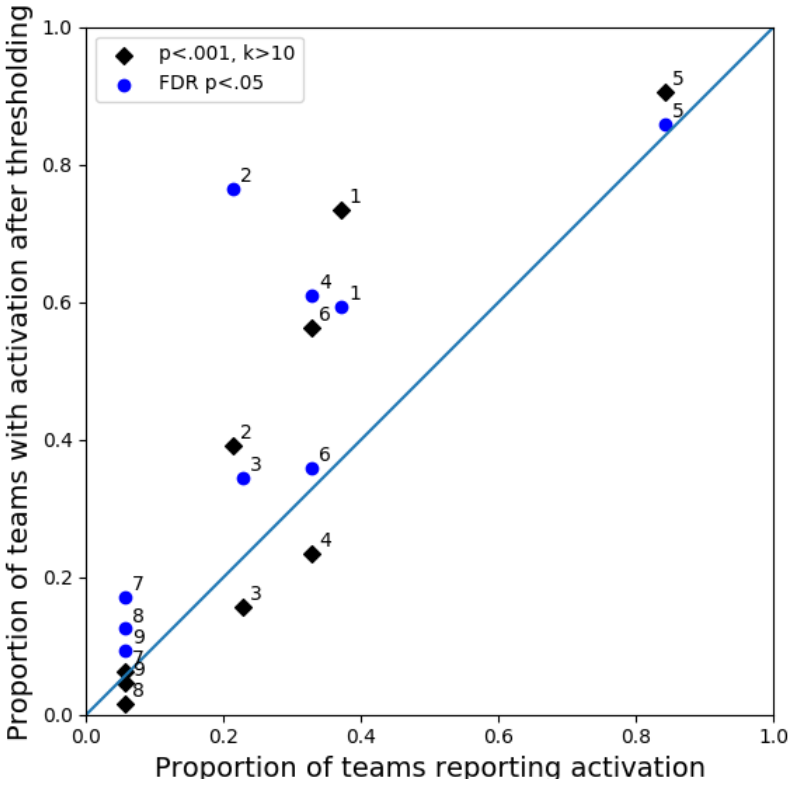


Extended Data Fig. 2 | Variability of whole-brain unthresholded maps for hypotheses 2 and 4–9. For each hypothesis, we present a heat map based on Spearman correlations between unthresholded statistical maps ($n = 64$), clustered according to their similarity, and the average of unthresholded images for each cluster (cluster colours in titles refer to colours in left margin of heat map). Column colours represent hypothesis decisions (green, yes; red, no) reported by the analysis teams; row colours denote cluster membership. Maps are thresholded at an uncorrected value of $z > 2$ for visualization. Unthresholded maps for hypotheses 2 and 4 are identical (as they both relate to the same contrast and group but different regions), and the colours represent reported results for hypothesis 2. For hypotheses 1 and 3, see Fig. 2.



Extended Data Fig. 3 | Variability and consensus of unthresholded statistical maps. $n = 64$. **a**, Maps of estimated between-team variability (tau) at each voxel for each hypothesis. **b**, Results of the image-based meta-analysis. A consensus analysis was performed on the unthresholded statistical maps to obtain a group statistical map for each hypothesis, accounting for the correlation between teams owing to the same underlying data (see Methods). Maps are presented for each hypothesis, showing voxels (in colour) in which the group statistic was significantly greater than zero after voxelwise correction for FDR ($P < 0.05$). Colour bar reflects statistical value (z) for the meta-analysis. Hypotheses 1 and 3, as well as hypotheses 2 and 4, share the same unthresholded maps, as they relate to the same contrast and group but different regions (see Extended Data Table 1). Images can be viewed at https://identifiers.org/neurovault.collection:6051.

a



b

Hypothesis	N voxels in ROI	Proportion of teams reporting activation	Proportion of teams with activation (p < 0:001, k > 10)	Proportion of teams with activation (FDR)	IBMA (n voxels in ROI)
1	3402	0.371	0.734	0.594	0
2	3402	0.214	0.391	0.766	7
3	173	0.229	0.156	0.344	0
4	173	0.329	0.234	0.609	7
5	3402	0.843	0.906	0.859	2101
6	3402	0.329	0.562	0.359	39
7	672	0.057	0.062	0.172	0
8	672	0.057	0.016	0.125	0
9	672	0.057	0.047	0.094	0

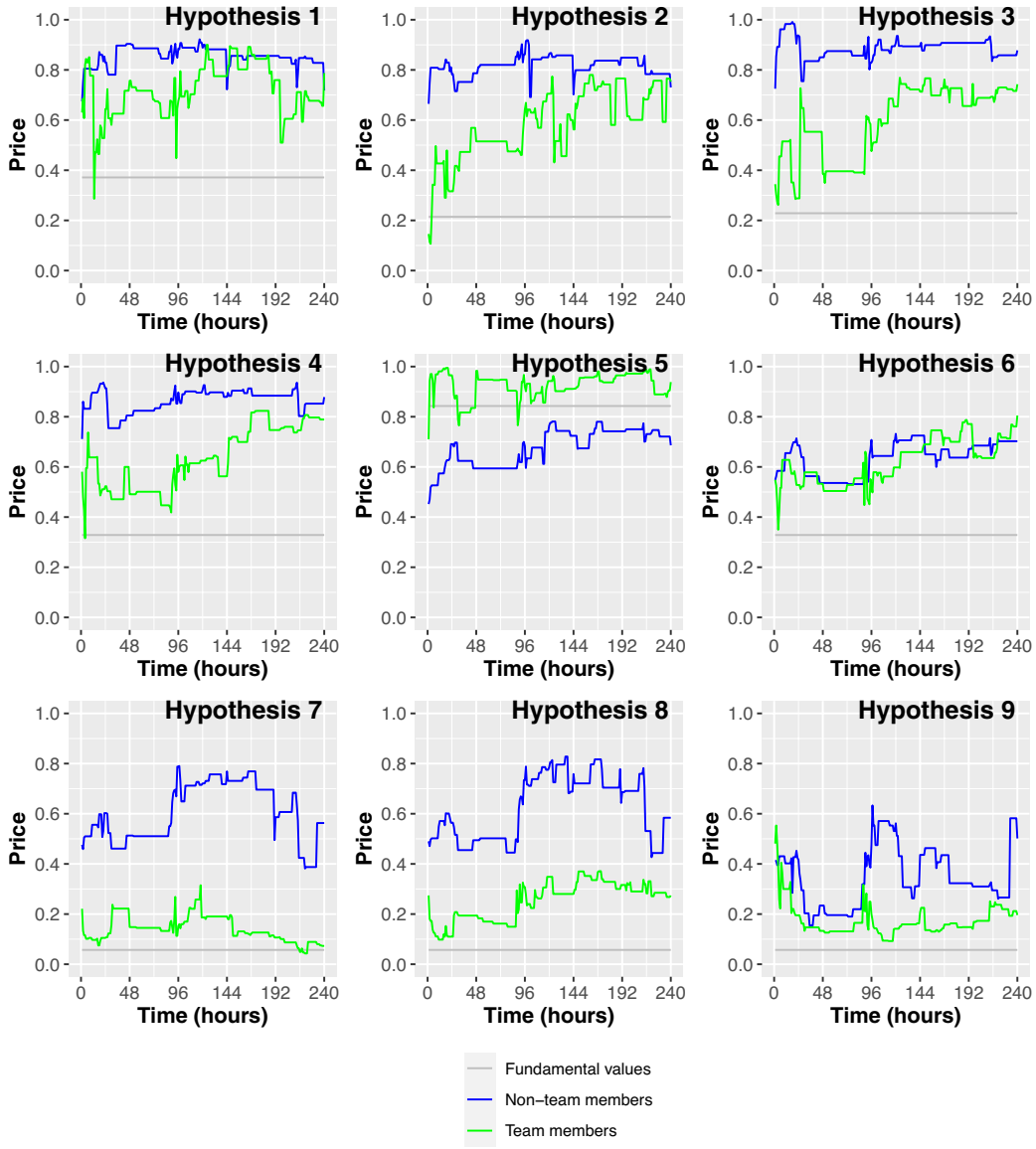
Extended Data Fig. 4 | Results of the consistent thresholding and ROI selection analysis. $n = 64$. **a**, Activation for each hypothesis as determined using consistent thresholding (black, $P < 0.001$ and cluster size (k) > 10 voxels; blue, FDR correction with $P < 0.05$) and ROI selection across teams (y axis), versus the actual proportion of teams reporting activation (x axis). Numbers next to each symbol represent the hypothesis number for each point. **b**, Results from re-thresholding of unthresholded maps, using either uncorrected values with the threshold ($P < 0.001$, $k > 10$) or FDR correction ($P_{FDR} < 5\%$) and common anatomical ROIs for each hypothesis. A team is recorded as having an activation if one or more significant voxels are found in the ROI. Results for image-based meta-analysis (IBMA) for each hypothesis are presented, also thresholded at $P_{FDR} < 5\%$.

a

Effect	Beta (full model)	t (full model)	p (full model)	Beta (no interaction)	t (no interaction)	p (no interaction)
Intercept	0.44	64.12	0.00	0.41	74.61	0.00
Time	0.00	3.38	0.00	0.00	12.48	0.00
Teams	-0.29	-29.50	0.00	-0.22	-45.35	0.00
Time X Teams	0.00	7.78	0.00			

Adjusted R-squared			0.35			0.34

b



Extended Data Fig. 5 | Prediction markets over time. $n = 240$ observations (10 days × 24 h). **a**, Panel regressions. The table summarizes the results of preregistered fixed-effects panel regressions of the absolute errors of the predictions (that is, the absolute deviation of the market price from the fundamental value) on an hourly basis (average price of all transactions within an hour) on time and prediction market indicators. Standard errors were computed using a robust estimator. **b**, Market prices for each of the nine hypotheses separated for the team members (green) and non-team members (blue) prediction markets. The figure shows the average prices of the prediction market per hour, separated for the two prediction markets, for the time the markets were open (10 days, that is, 240 h). The grey line indicates the actual share of the analysis teams that reported a significant result for the hypothesis (that is, the fundamental value).

Extended Data Table 1 | Hypotheses and results

	Hypothesis description	Fraction of teams reporting a significant result	Median confidence level	Median similarity estimation
#1	Positive parametric effect of gains in the vmPFC (equal indifference group)	0.371	7 (2)	7 (1.5)
#2	Positive parametric effect of gains in the vmPFC (equal range group)	0.214	7 (1.5)	7 (1)
#3	Positive parametric effect of gains in the ventral striatum (equal indifference group)	0.229	6 (1)	7 (1)
#4	Positive parametric effect of gains in the ventral striatum (equal range group)	0.329	6 (1)	7 (1)
#5	Negative parametric effect of losses in the vmPFC (equal indifference group)	0.843	8 (1)	8 (1)
#6	Negative parametric effect of losses in the vmPFC (equal range group)	0.329	7 (1)	7 (1)
#7	Positive parametric effect of losses in the amygdala (equal indifference group)	0.057	7 (1)	8 (1)
#8	Positive parametric effect of losses in the amygdala (equal range group)	0.057	7 (1)	8 (1)
#9	Greater positive response to losses in amygdala (equal range group vs. equal indifference group)	0.057	6 (1)	7 (1)

Each hypothesis is described along with the fraction of teams that reported a whole-brain-corrected significant result (out of $n = 70$ teams) and two measures reported by the analysis teams for the specific hypothesis: (1) How confident are you about this result? (2) How similar do you think your result is to the other analysis teams? Both of these ordinal measures are rated on a scale of 1–10, and the median values are presented together with the median absolute deviation in brackets. vmPFC, ventromedial prefrontal cortex. See Supplementary Information for analysis of the confidence level and similarity estimation.

Extended Data Table 2 | Results submitted by analysis teams

Team ID	H1	H2	H3	H4	H5	H6	H7	H8	H9	Est. smoothing	Package	fMRIPrep	Testing	Movement
08MQ	8	6	8	6	7	7	7	7	6	13.14	FSL	No	Non-parametric	Yes
0C7Q	7	7	8	8	8	7	10	10	9	8.68	Other	Yes	Non-parametric	Yes
0ED6	7	9	8	7	8	8	9	9	6	7.86	SPM	No	Parametric	Yes
0H5E	4	7	7	6	8	5	8	7	1	14.17	SPM	No	Parametric	No
0I4U	4	7	6	8	9	9	9	9	9	8.69	SPM	No	Parametric	Yes
0JO0	7	5	5	5	5	5	5	5	5	8.12	Other	Yes	Parametric	Yes
16IN	8	7	6	6	8	7	8	6	6		Other	Yes	Other	No
1K0E	7	9	6	6	8	7	7	6	9		Other	No	Non-parametric	Yes
1KB2	6	6	8	8	5	5	8	8	7	13.06	FSL	No	Parametric	Yes
1P0Y	8	8	1	1	8	8	5	5	5	9.13	SPM	No	Parametric	No
27SS	4	6	7	7	7	7	6	8	4	11.37	AFNI	No	Parametric	Yes
2T6S	8	9	6	6	10	9	7	8	10	14.93	SPM	Yes	Parametric	Yes
2T7P	8	8	8	8	8	8	8	8	8	7.66	Other	No	Other	Yes
3C6G	6	7	7	5	8	8	8	8	8	14.26	SPM	No	Parametric	Yes
3PQ2	9	8	7	7	7	8	8	8	7	5.79	FSL	No	Parametric	Yes
3TR7	2	2	3	4	8	5	8	6	5	17.4	SPM	Yes	Parametric	Yes
43FJ	3	3	5	5	10	10	10	10	10	10.66	FSL	No	Parametric	Yes
46CD	9	8	5	8	9	8	9	9	5	10.92	Other	No	Parametric	Yes
4SZ2	7	5	6	6	9	9	7	8	7	6.65	FSL	Yes	Parametric	No
4TQ6	7	9	10	9	7	8	10	10	9	14.88	FSL	Yes	Non-parametric	No
50GV	10	10	10	10	10	10	10	10	10	10.26	FSL	Yes	Parametric	No
51PW	8	8	8	8	8	8	6	6	7	11.15	FSL	Yes	Parametric	Yes
5G9K	7	7	7	7	7	7	7	7	7		SPM	Yes	Parametric	Yes
6FH5	9	2	8	8	10	8	8	9	9	12.22	SPM	No	Parametric	Yes
6VV2	8	8	8	6	9	7	8	7	6	7.2	AFNI	No	Parametric	Yes
80GC	9	9	8	4	3	9	6	5	4	4.02	AFNI	Yes	Parametric	Yes
94GU	8	8	8	8	8	8	8	8	8	11.19	SPM	No	Parametric	Yes
98BT	9	7	7	8	9	7	8	8	8	11.48	SPM	No	Parametric	Yes
9Q6R	10	10	10	10	10	10	8	8	8	10.28	FSL	No	Parametric	Yes
9T8E	5	5	5	5	5	5	5	5	4	9.85	SPM	Yes	Non-parametric	Yes
9U7M	7	9	9	9	9	7	9	7	7	14.78	Other	No	Parametric	Yes
AO86	7	7	7	7	7	7	7	7	7	7.49	Other	Yes	Non-parametric	Yes
B23O	6	6	7	7	8	7	6	6	8	3.32	FSL	Yes	Non-parametric	No
B5I6	10	10	5	5	10	6	8	7	6	9.84	FSL	Yes	Non-parametric	Yes
C22U	8	7	5	8	9	8	8	8	8	11.16	FSL	No	Parametric	No
C88N	7	8	7	4	9	7	8	8	6	11.62	SPM	Yes	Parametric	No
DC61	5	1	5	2	9	5	5	5	5	9.58	SPM	Yes	Parametric	Yes
E3B6	3	7	6	6	8	8	7	7	7	12.8	SPM	Yes	Parametric	Yes
E6R3	5	5	7	3	4	4	7	7	7	9.28	Other	Yes	Other	Yes
I07H	3	3	3	3	9	9	9	9	9	5.59	Other	Yes	Non-parametric	No
I52Y	8	8	8	8	8	8	8	8	8	11.42	FSL	No	Non-parametric	Yes
I9D6	7	7	7	7	1	7	7	6	7	6.21	AFNI	No	Parametric	Yes
IZ20	7	7	7	7	7	7	7	6	6	21.28	Other	No	Parametric	No
J7F9	9	8	9	7	9	7	9	9	9	14.88	SPM	Yes	Parametric	Yes
K9P0	10	10	10	5	10	8	9	9	10	8.05	AFNI	Yes	Parametric	Yes
L1A8	8	5	7	7	8	8	3	8	3		SPM	No	Parametric	Yes
L3V8	9	9	9	9	9	9	9	9	9	14.74	SPM	No	Parametric	No
L7J7	10	9	9	5	8	8	8	9	8	11.76	SPM	Yes	Parametric	Yes
L9G5	5	4	4	6	10	10	9	9	7	7.22	FSL	No	Parametric	No
O03M	3	8	8	2	8	7	7	7	7	3.47	AFNI	Yes	Non-parametric	Yes
O21U	8	8	8	8	8	8	8	8	8	8.26	FSL	Yes	Parametric	Yes
O6R6	8	8	8	8	8	8	8	8	8	3.06	FSL	Yes	Non-parametric	No
P5F3	3	5	7	7	4	4	6	6	7	12.94	FSL	No	Parametric	Yes
Q58J	9	9	9	9	9	9	9	9	9	16.24	FSL	No	Parametric	No
Q6O0	7	8	8	9	9	8	8	6	7	14.58	SPM	Yes	Parametric	Yes
R42Q	5	5	6	6	6	6	7	8	8	12.73	Other	No	Parametric	Yes
R5K7	6	8	8	7	9	7	8	8	7	12.06	SPM	No	Parametric	Yes
R7D1	4	7	5	5	9	5	8	9	8	8.93	Other	Yes	Non-parametric	Yes
R9K3	5	3	2	5	8	5	3	4	5	11.77	SPM	Yes	Parametric	Yes
SM54	5	9	5	8	8	6	8	8	8	7.05	Other	Yes	Parametric	Yes
T54A	5	9	2	6	9	9	5	5	5	12.28	FSL	Yes	Non-parametric	No
U26C	8	8	8	8	10	8	8	8	9	10.38	SPM	Yes	Parametric	Yes
UI76	10	6	10	10	10	6	10	10	5	6.6	AFNI	Yes	Parametric	Yes
UK24	4	4	4	4	4	4	4	4	4	10.76	SPM	No	Parametric	No
V55J	4	5	7	7	4	7	5	7	7	12.85	SPM	No	Parametric	No
VG39	6	7	8	8	10	7	9	6	5		SPM	Yes	Parametric	No
X19V	6	7	8	5	9	6	9	9	9	8.48	FSL	Yes	Parametric	Yes
X1Y5	6	6	7	7	8	6	8	8	8	8.69	Other	Yes	Non-parametric	Yes
X1Z4	8	6	4	4	9	5	4	4	4		Other	No	Non-parametric	Yes
XU70	4	5	8	9	9	9	6	8	8	7.17	FSL	No	Parametric	Yes

For each team, the left section of the table represents the reported binary decision (green, yes; red, no) and how confident they were in their result (from 1 (not at all confident) to 10 (extremely confident)) for each hypothesis (H1–H9). The right section displays the information included for each team in the statistical model for hypothesis decisions. Estimated (est.) smoothing values represent full width at half-maximum (FWHM); teams with a blank value were excluded from further analysis. Note that three teams changed their decisions after the end of the project: team L3V8 changed its decision for hypothesis 6 from yes to no; team VG39 changed its decisions for hypotheses 3, 4 and 5 from yes to no; and team U26C changed its decision for hypothesis 5 from yes to no. Results throughout the paper and in this table reflect the final results as they were reported at the end of the project (that is, before this change), as prediction markets were based on those results.

Extended Data Table 3 | Data links and analysis-related tables

a

Team ID	Collection	Team ID	Collection
08MQ	4953	C88N	4812
0C7Q	5652	DC61	4963
0ED6	4994	E3B6	4782
0H5E	4936	E6R3	4959
0I4U	4938	I07H	5001
0JO0	4807	I52Y	4933
16IN	4927	I9D6	4978
1K0E	4974	IZ20	4979
1KB2	4945	J7F9	4949
1P0Y	5649	K9P0	4961
27SS	4975	L1A8	5680
2T6S	4881	L3V8	4888
2T7P	4917	L7J7	4866
3C6G	4772	L9G5	5173
3PQ2	4904	O03M	4972
3TR7	4966	O21U	4779
43FJ	4824	O6R6	4907
46CD	5637	P5F3	4967
4SZ2	5665	Q58J	5164
4TQ6	4869	Q6O0	4968
50GV	4735	R42Q	5619
51PW	5167	R5K7	4950
5G9K	4920	R7D1	4954
6FH5	5663	R9K3	4802
6VV2	4883	SM54	5675
80GC	4891	T54A	4876
94GU	5626	U26C	4820
98BT	4988	UI76	4821
9Q6R	4765	UK24	4908
9T8E	4870	V55J	4919
9U7M	4965	VG39	5496
AO86	4932	X19V	4947
B23O	4984	X1Y5	4898
B5I6	4941	X1Z4	4951
C22U	5653	XU70	4990

b

Team ID	Exclusion reason	Unthresholded maps excluded	Thresholded maps excluded
1K0E	Used surface-based analysis (only provided data for cortical ribbon)	X	X
L1A8	Not in MNI standard space	X	X
VG39	Performed small volume corrected instead of whole-brain analysis	X	X
X1Z4	Used surface-based analysis (only provided data for cortical ribbon)	X	X
16IN	Values in the unthresholded images are not z / t stats	X	
5G9K	Values in the unthresholded images are not z / t stats	X	
2T7P	Used a method which does not create thresholded images (and are therefore not included in the analyses of the thresholded images)		X

c

Effects	Chi-squared	P value	Delta R2
Hypothesis	185.390	0.000	0.350
Estimated smoothness	13.210	0.000	0.040
Used fMRIPprep data	2.270	0.132	0.010
Software package	13.450	0.004	0.040
Multiple correction method	7.500	0.024	0.020
Movement modeling	1.160	0.281	0.000

d

Team ID	Link to shared analysis codes
16IN	https://github.com/jennyrieck/NARPS
2T7P	https://osf.io/3b57r
E3B6	doi.org/10.5281/zenodo.3518407
Q58J	https://github.com/amrka/NARPS_Q58J

a, Numbers of public NeuroVault collections of all analysis teams https://neurovault.org/collections/. **b**, Descriptions of teams that were excluded from the analyses of statistical maps. **c**, Summary of mixed-effects logistic regression modelling of decision outcomes (n = 64 per hypothesis) as a function of different factors including the hypothesis (1–9) and various aspects of statistical modelling (for modelling details see https://github.com/poldrack/narps/blob/master/ImageAnalyses/DecisionAnalysis.Rmd). **d**, Links to shared analysis code of some of the analysis teams.

Extended Data Table 4 | Variability of statistical maps across teams

a

Hypothesis	Minimum sig. voxels	Maximum sig. voxels	Median sig. voxels	N empty images
1	0	118181	1940	8
2	0	135583	8120	2
3	0	118181	1940	8
4	0	135583	8120	3
5	0	76569	6527	11
6	0	72732	167	25
7	0	147087	9383	8
8	0	129979	475	16
9	0	49062	266	29

b

Hypothesis	Correlation (mean)	Cluster1		Cluster2		Cluster3	
		Correlation	Cluster size	Correlation	Cluster size	Correlation	Cluster size
1+3	0.394	0.670	50	0.680	7	0.095	7
2+4	0.521	0.736	43	0.253	14	0.659	7
5	0.485	0.777	41	0.329	20	0.342	3
6	0.259	0.442	47	0.442	12	0.156	5
7	0.487	0.851	31	0.466	25	0.049	8
8	0.302	0.593	36	0.256	23	-0.044	5
9	0.205	0.561	47	0.568	8	0.106	9

a, Variability in the number of significantly (sig.) activated voxels reported across teams (n = 65 teams). **b**, Mean Spearman correlation between the unthresholded statistical maps for all pairs of teams and separately for pairs of teams within each cluster, for each hypothesis (n = 64 teams).

Extended Data Table 5 | Results of prediction markets and additional data

a

Hypothesis	FV	CI	Non-teams market prediction	Teams market prediction
1	0.37	[0.26-0.48]	0.727 *	0.814 *
2	0.21	[0.12-0.31]	0.73 *	0.753 *
3	0.23	[0.13-0.33]	0.881 *	0.743 *
4	0.33	[0.22-0.44]	0.882 *	0.789 *
5	0.84	[0.76-0.93]	0.686 *	0.952 *
6	0.33	[0.22-0.44]	0.685 *	0.805 *
7	0.06	[0.00-0.11]	0.563 *	0.073
8	0.06	[0.00-0.11]	0.584 *	0.274 *
9	0.06	[0.00-0.11]	0.476 *	0.188 *

b

Hypothesis	1	2	3	4	5	6	7	8	9
Spearman rho	0.58	0.56	0.58	0.64	0.47	0.74	0.23	0.37	0.31
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.01	0.02
Share of consistent holdings	0.71	0.68	0.70	0.80	0.89	0.74	0.80	0.80	0.75
Z (signed rank test)	3.40	2.78	2.82	4.24	6.81	3.24	4.34	4.34	3.64
p-value (signed rank test)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average holdings if consistent	5.61	21.14	25.80	13.11	-115.50	7.31	34.61	24.23	23.54
Average holdings if inconsistent	1.04	-6.90	-8.03	0.03	18.26	1.58	-14.63	-8.29	-11.61

c

Hypothesis	Tokens invested (Non-teams)	Volume (Non-teams)	# Traders (Non-teams)	# Transactions (Non-teams)	Tokens invested (Teams)	Volume (Teams)	# Traders (Teams)	# Transactions (Teams)
1	8.568	20.175	55	139	12.643	25.671	64	213
2	10.51	22.544	53	98	11.632	22.908	58	171
3	12.818	24.709	58	132	7.773	15.837	52	141
4	11.134	20.397	49	112	8.126	15.479	52	127
5	6.873	14.636	38	71	14.48	30.76	76	244
6	6.806	12.663	35	72	8.097	16.676	46	134
7	7.99	15.209	41	98	7.131	15.864	52	160
8	8.791	19.072	45	91	7.085	14.598	52	141
9	10.427	21.118	50	131	9.506	18.812	56	178

a, Summary of the prediction market results. FV refers to the fundamental value, that is, the actual fraction of teams (out of $n = 70$ teams) that reported significant results for the hypothesis. CI refers to the 95% confidence interval corresponding to the fundamental value (estimated with a normal approximation to the binomial distribution). Values marked with an asterisk are not within the corresponding 95% CI. **b**, Consistency of traders' holdings and team results. The top two rows show two-sided Spearman rank correlations between traders' final holdings and the binary result reported by their team, and the corresponding *P* value for each hypothesis. The bottom five rows show the share of traders' holdings that are consistent with the results reported by their team. Consistent refers to positive (negative) holdings if the team reported a significant (non-significant) result; *z* and *P* values refer to Wilcoxon signed-rank tests for the share of consistent holdings being equal to 0.5; and average holdings if (in)consistent refer to the mean final holdings, separated for consistent and inconsistent traders. **c**, Additional data for each of the nine hypotheses. Tokens invested indicates the average number of tokens invested per transaction; volume refers to the mean number of shares bought or sold per transaction; # traders refers to the number of traders who bought or sold shares of the particular asset at least once; and # transactions describes the overall number of transactions recorded.

nature research

Corresponding author(s): Tom Schonberg

Last updated by author(s): Mar 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	A full description of the experimental procedures, validations and the fMRI dataset is available in a Data Descriptor (https://doi.org/10.1038/s41597-019-0113-7). Code used for fMRI data collection are available at https://github.com/rotemb9/NARPS_scientific_data.
Data analysis	Fully reproducible code for the analyses of the analysis teams' submitted results and statistical maps, as well as the prediction markets, are available at DOI: 10.5281/zenodo.3709273. The full list of software and versions used within the code are available in the dockerfile: https://github.com/poldrack/narps/blob/master/Dockerfile

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The fMRI dataset is openly available via OpenNeuro at DOI:10.18112/openneuro.ds001734.v1.0.4. Additional data are included with the analyses code at DOI:10.5281/zenodo.3709273

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative
Research sample	The fMRI dataset included nueroimaging and behavioral data of 108 participants. Demographic information of the participants can be found at DOI:10.18112/openneuro.ds001734.v1.0.4. 70 analysis teams analyzed the dataset. 96 “team members” and 91 “non-team members” signed up to participate in the prediction markets. N = 83 “team members” and N = 65 “non-team members” actively participated in the markets. Members of the analysis teams and traders in the predictions market were researchers in the field from around the world.
Sampling strategy	Relevant information for the fMRI dataset is available at the Data Descriptor (https://doi.org/10.1038/s41597-019-0113-7). With regard to the number of analysis teams and traders in the prediction markets, we aimed to recruit as many as possible within the time frame.
Data collection	Relevant information for the fMRI dataset is available at the Data Descriptor (https://doi.org/10.1038/s41597-019-0113-7). Shortly, data was collected using MRI scanner and computers.
Timing	The fMRI dataset was collected between November 2017 and May 2018. Analysis teams were recruited and analyzed the data between November 2018 and March 2019. The prediction markets were open between May 2nd to May 12th 2019.
Data exclusions	One team was excluded from all analyses since their reported results were not based on a whole-brain analysis as instructed. Of the remaining 69 teams, thresholded maps of 65 teams and unthresholded (z / t) maps of 64 teams were included in the analyses (see Extended Data Table 3b for detailed reasons for exclusion of the other teams).
Non-participation	12 out of the 82 analysis teams that signed the non-disclosure form and were provided with access to the data did not submit their results by the deadline. 13 traders in the "team members" and 26 traders in the "non-team members" prediction markets registered but did not actively participate in the prediction markets.
Randomization	fMRI dataset- participants were pseudo-randomly (alternately) assigned to one of two experimental conditions (Equal Indifference or Equal Range). Analysis teams were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	fMRI dataset- demographic information of the participants can be found at DOI:10.18112/openneuro.ds001734.v1.0.4. 108 participants where included in the dataset: 54 in the Equal Indifference group (30 females, mean age = 26.06 years, SD age = 3.02 years) and 54 in the Equal Range group (30 females, mean age = 25.04 years, SD age = 3.99 years). All participants were right-handed, had normal or corrected-to-normal vision and reported no history of psychiatric or neurologic diagnoses, or use any medications that would interfere with the experiment.

Recruitment	Analysis teams were recruited via social media, mainly Twitter and Facebook, as well as during the 2018 annual meeting of The Society for Neuroeconomics. Prediction market traders were recruited via social media (mainly Facebook and Twitter) and e-mails. This recruitment method may increase the chances of specific researchers to participate in an analysis team or in the prediction markets, for example researchers that are more active in social media or attended the 2018 meeting of The Society for Neuroeconomics. Researchers who advocate for replication attempts and "open science" practices may also be more inclined to join such study. However, our results strongly suggest that they were not biased. For example, the fact that several hypotheses were only affirmed by roughly 5% of teams, while Hypothesis #5 was affirmed by 84% of teams, suggests that there was no overall bias towards either affirmation or rejection of hypotheses. In addition, each of the 70 analysis teams chose to use a different analysis pipeline, which suggests evidence against a potential bias in methods used by the specific analysis teams that joined the study. With regard to the prediction markets, traders that were exposed to the recruitment ads on social media may be biased with regard to their predictions, but as there is a debate in the published literature regarding most of the hypotheses included in our study, we do not have a specific reason to assume such bias.
Ethics oversight	MRI data collection was approved by the Helsinki committee at Sheba Tel Hashomer Medical Center and the ethics committee at Tel Aviv University, and all participants gave written informed consent (as described in the Scientific Data Descriptor of this dataset). The Board for Ethical Questions in Science at the University of Innsbruck approved the data collection in regards of the prediction markets, and certified that the project is in correspondence with all requirements of the ethical principles and the guidelines of good scientific practice. The Stanford University IRB determined that the analysis of the submitted team results did not meet the definition of human subject research, and thus no further IRB review was required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Task
Design specifications	The fMRI dataset was published in a Data Descriptor (https://doi.org/10.1038/s41597-019-0113-7)
Behavioral performance measures	The fMRI dataset was published in a Data Descriptor (https://doi.org/10.1038/s41597-019-0113-7)

Acquisition

Imaging type(s)	functional and structural
Field strength	3T
Sequence & imaging parameters	Imaging data were acquired using a 3T Siemens Prisma MRI scanner with a 64-channel head coil, at the Strauss Imaging Center on the campus of Tel Aviv University. Functional data during the mixed gambles task were acquired using T2*-weighted echo-planar imaging sequence with multi-band acceleration factor of 4 and parallel imaging factor (iPAT) of 2, TR=1000ms, TE=30ms, flip angle=68 degrees, field of view (FOV)=212×212mm, in plane resolution of 2×2mm 30 degrees off the anterior commissure-posterior commissure line to reduce the frontal signal dropout27, slice thickness of 2mm, 64 slices and a gap of 0.4mm between slices to cover the entire brain. For each functional run, we acquired 453 volumes.
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Each team performed their own preprocessing. Raw data and data preprocessed with fMRIprep v. 1.1.4 were shared with the teams.
Normalization	Each team performed their own preprocessing. Raw data and data preprocessed with fMRIprep v. 1.1.4 were shared with the teams.
Normalization template	Each team performed their own preprocessing. Raw data and data preprocessed with fMRIprep v. 1.1.4 were shared with the teams.
Noise and artifact removal	Each team performed their own preprocessing. Raw data and data preprocessed with fMRIprep v. 1.1.4 were shared with the teams.
Volume censoring	Each team performed their own preprocessing. Raw data and data preprocessed with fMRIprep v. 1.1.4 were shared with the teams.

Statistical modeling & inference

Model type and settings	Each team performed their own analysis.
Effect(s) tested	Each team performed their own analysis.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	Each team performed their own analysis.
Correction	Each team performed their own analysis.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis