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

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[37,9], Luca Cecchetti[38], Edna C. Cieslik[35,36], Zachary J. Cole[39], Olivier Collignon[19,40], Robert W. Cox[41], William A. Cunningham[42], Stefan Czoschke[43],Kamalaker Dadi[44], Charles P. Davis[45,46,47], Alberto De Luca[48], Mauricio R. Delgado[25], Lysia Demetriou[49,50], Jeffrey B. Dennison[51], Xin Di[26,52], Erin W. Dickie[53,54], Ekaterina Dobryakova[55], Claire L. Donnat[56], Juergen Dukart[35,36], Niall W. Duncan[57,58], Joke Durnez[59], Amr Eed[60], Simon B. Eickhoff[35,36], Andrew Erhart[34],Laura Fontanesi[61], G. Matthew Fricke[62], Shiguang Fu[63,64], Adriana Galván[65], Remi Gau[19], Sarah Genon[35,36], Tristan Glatard[66], Enrico Glerean[67], Jelle J. Goeman[68], Sergej A. E. Golowin[57], Carlos González-García[69], Krzysztof J. Gorgolewski[70], Cheryl L. Grady[71], Mikella A. Green[9,37], João F. Guassi Moreira[65], Olivia Guest[28,72], Shabnam Hakimi[9], J. Paul Hamilton[73], Roeland Hancock[46,47], Giacomo Handjaras[38], Bronson B. Harry[74], Colin Hawco[75], Peer Herholz[76], Gabrielle Herman[75], Stephan Heunis[77,78], Felix Hoffstaedter[35,36], Jeremy Hogeveen[79], Susan Holmes[80], Chuan-Peng Hu[81], Scott A. Huettel[82], Matthew E. Hughes[83,84], Vittorio Iacovella[12], Alexandru D. Iordan[85], Peder M. Isager[86], Ayse I. Isik[87], Andrew Jahn[88], Matthew R. Johnson[39,89], Tom Johnstone[90], Michael J. E. Joseph[91], Anthony C. Juliano[92], Joseph W. Kable[93,94], Michalis Kassinopoulos[95], Cemal Koba[38], Xiang-Zhen Kong[96], Timothy R. Koscik[97], Nuri Erkut Kucukboyaci[55,98], Brice A. Kuhl[99], Sebastian Kupek[100], Angela R. Laird[101], Claus Lamm[102,103], Robert Langner[35,36], Nina Lauharatanahirun[104,105], Hongmi Lee[106], Sangil Lee[93], Alexander Leemans[48], Andrea Leo[38], Elise Lesage[32], Flora Li[107,108], Monica Y.C. Li[45,46,47,109], Phui Cheng Lim[39,89], Evan N. Lintz[39], Schuyler W. Liphardt[110], Annabel B. Losecaat Vermeer[102], Bradley C. Love[28,111], Michael L. Mack[42], Norberto Malpica[112], Theo Marins[29], Camille Maumet[113], Kelsey McDonald[37], Joseph T. McGuire[15,16], Helena Melero[112,114,115], Adriana S. Méndez Leal[65], Benjamin Meyer[116,117], Kristin N. Meyer[118], Glad Mihai[119,120], Georgios D. Mitsis[121], Jorge Moll[29], Dylan M. Nielson[122], Gustav Nilsonne[123,124], Michael P. Notter[125], Emanuele Olivetti[11,12], Adrian I. Onicas[38], Paolo Papale[38,126], Kaustubh R. Patil[35,36], Jonathan E. Peelle[14], Alexandre Pérez[76], Doris Pischedda[127,128,129], Jean-Baptiste Poline[76,130], Yanina Prystauka[45,46,47], Shruti Ray[131], Patricia A. Reuter-Lorenz[85], Richard C. Reynolds[132], Emiliano Ricciardi[38], Jenny R. Rieck[71], Anais M. Rodriguez-Thompson[118], Anthony Romyn[133], Taylor Salo[134], Gregory R. Samanez-Larkin[9,37], Emilio Sanz-Morales[112], Margaret L. Schlichting[42], Douglas H. Schultz[39,89], Qiang Shen[63,64], Margaret A. Sheridan[135], Jennifer A. Silvers[65], Kenny Skagerlund[136,137], Alec Smith[138], David V. Smith[51], Peter Sokol-Hessner[34], Simon R. Steinkamp[139], Sarah M. Tashjian[65], Bertrand Thirion[140], John N. Thorp[141], Gustav Tinghög[142,143], Loreen Tisdall[144,145], Steven H. Tompson[104], Claudio Toro-Serey[15,16], Juan Jesus Torre Tresols[140], Leonardo Tozzi[146], Vuong Truong[57,58], Luca Turella[12], Anna E. van 't Veer[147], Tom Verguts[32], Jean M. Vettel[148,149,150], Sagana Vijayarajah[42], Khoi Vo[151,152], Matthew B. Wall[153,154,155], Wouter D. Weeda[147], Susanne Weis[35,36], David J. White[156], David Wisniewski[32], Alba Xifra-Porxas[95], Emily A. Yearling[45,46,47], Sangsuk Yoon[157], Rui Yuan[158], Kenneth S.L. Yuen[81,116], Lei Zhang[102], Xu Zhang[46,47,159], Joshua E. Zosky[39,89], Thomas E. Nichols[31,*], Russell A. Poldrack[145,*], Tom Schonberg[1,2,*]

[*] Corresponding authors

[1]Sagol School of Neuroscience, Tel Aviv University, Israel; [2]Faculty of Life Sciences, Department of Neurobiology, Tel Aviv University, 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, 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 Saint Louis, Saint 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, Toronto, Canada; [22]Department of Psychology, Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands; [23]Department of Psychology, Scene Grammar Lab, 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, Florida, USA; [31]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of 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 (INM-7: Brain and Behaviour), Research Centre Juelich, Juelich, Germany; [36]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Duesseldorf, Duesseldorf, 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]Center for Mind and Brain Science, University of Trento, Trento, Italy; [41]NIMH/NIH, Bethesda, MD, USA; [42]Department of Psychology, University of Toronto, Toronto, ON, Canada; [43]Institute of Medical Psychology, Goethe University, Frankfurt am Main, Germany; [44] Inria, CEA, Université Paris-Saclay, Palaiseau, 91120, France; [45]Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA; [46]Brain Imaging Research Center, University of Connecticut, Storrs, CT, USA; [47]Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT, USA; [48]PROVIDI Lab, Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands; [49]Section of Endocrinology & Investigative Medicine, Faculty of Medicine, Imperial College London, London, UK; [50]Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK; [51]Department of Psychology, Temple University, Philadelphia, PA, USA; [52]School of Life Sciences and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan, China; [53]Krembil Centre for Neuroinformatics, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada; [54]Department of Psychiatry, University of Toronto, Canada; [55]Center for Traumatic Brain Injury Research, Kessler Foundation, East Hanover, NJ, USA; [56]Department of Statistics, Stanford University, Stanford, CA, USA; [57]Graduate Institute of Mind, Brain and Consciousness, Taipei Medical University, Taipei, Taiwan; [58]Brain and Consciousness Research Centre, TMU-ShuangHo Hospital, New Taipei City, Taiwan; [59]Department of Psychology and Stanford Center for Reproducible Neuroscience, Stanford University, Stanford, 94305, California, USA; [60]Instituto de Neurociencias, CSIC-UMH, Spain; [61]Faculty of Psychology, University of Basel, Basel, Switzerland; [62]Computer Science Department, University of New Mexico, Albuquerque, NM, USA; [63]School of Management, Zhejiang University of Technology, Hangzhou, China; [64]Institute of Neuromanagement, Zhejiang University of Technology, Hangzhou, China; [65]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA; [66]Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada; [67]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland; [68]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands; [69]Department of Experimental Psychology, Ghent University, Belgium; [70]Department of Psychology, Stanford University, CA, USA; [71]Rotman Research Institute, Baycrest Health Sciences Centre, Toronto, Ontario, Canada;

[72]Research Centre on Interactive Media, Smart Systems and Emerging Technologies - RISE, Nicosia, Cyprus; [73]Center for Social and Affective Neuroscience, Department of Biomedical and Clinical Sciences, Linköping University, Sweden; [74]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, NSW, Australia; [75]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada; [76]McConnell Brain Imaging Centre, The Neuro (Montreal Neurological Institute-Hospital), Faculty of Medicine, McGill University, Montreal, QC, Canada; [77]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands; [78]Department of Research and Development, Epilepsy Centre Kempenhaeghe, Heeze, The Netherlands; [79]Department of Psychology & Psychology Clinical Neuroscience Center, University of New Mexico, Albuquerque, NM, USA; [80]Statistics Department, Stanford University, Stanford, CA, USA; [81]Leibniz-institut für Resilienzforschung (LIR), Mainz, Germany; [82]The Department of Psychology and Neuroscience, Duke University, Durham, NC, USA; [83]School of Health Sciences, Swinburne University of Technology, Hawthorn, Australia; [84]Australian National Imaging Facility (NIF), Australia; [85]Department of Psychology, University of Michigan, Ann Arbor, MI, USA; [86]Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands; [87]Neuroscience Department, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany; [88]fMRI Laboratory, University of Michigan, Ann Arbor, MI, USA; [89]Center for Brain, Biology and Behavior, University of Nebraska-Lincoln, Lincoln, NE, USA; [90]School of Health Sciences, Swinburne University of Technology, Hawthorn, VIC, Australia; [91]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada; [92]Center for Neuropsychology and Neuroscience Research, Kessler Foundation, East Hanover, NJ, USA; [93]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA; [94]MindCORE, University of Pennsylvania, Philadelphia, PA, USA; [95]Graduate Program in Biological and Biomedical Engineering, McGill University, Montreal, QC, Canada; [96]Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; [97]University of Iowa Carver College of Medicine, Department of Psychiatry, Iowa City, IA, USA; [98]Department of PM&R, Rutgers New Jersey Medical School, Newark, NJ; [99]Department of Psychology, University of Oregon, Eugene, OR, USA; [100]Faculty of Economics and Statistics, University of Innsbruck, Innsbruck, Austria; [101]Department of Physics, Florida International University, Miami, Florida, USA; [102]Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria; [103]Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria; [104]U.S. CCDC Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, MD, USA; [105]University of Pennsylvania, Annenberg School for Communication, Philadelphia, PA, USA; [106]The Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA; [107]Fralin Biomedical Research Institute, Roanoke, VA, USA; [108]Economics Experimental Lab, Nanjing Audit University, Nanjing, China; [109]Haskins Laboratories, New Haven, CT, USA; [110]Biology Department, University of New Mexico, Albuquerque, NM, USA; [111]The Alan Turing Institute, London, UK; [112]Laboratorio de Análisis de Imagen Médica y Biometría (LAIMBIO), Universidad Rey Juan Carlos, Madrid, Spain; [113]Inria, Univ Rennes, CNRS, Inserm, IRISA UMR 6074, Empenn ERL U 1228, Rennes, France; [114]Departamento de Psicobiología, División de Psicología, CES Cardenal Cisneros, Madrid, Spain; [115]Northeastern University Biomedical Imaging Center, Northeastern University, Boston, MA, USA; [16]Neuroimaging Center (NIC), Focus Program Translational Neurosciences (FTN), Johannes Gutenberg University Medical Center Mainz, Germany; [117]Leibniz-Institut für Resilienzforschung (LIR), Mainz, Germany; [118]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [119]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany; [120]Technische Universität Dresden, Germany; [121]Department of Bioengineering, McGill University, QC, Canada; [122]Data Science and Sharing Team, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA; [123]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden; [124]Department of Psychology, Stockholm University, Stockholm, Sweden; [125]The Laboratory for Investigative Neurophysiology (The LINE), Department of Radiology, University Hospital Center and University of Lausanne, Switzerland; [126]Department of Vision & Cognition, Netherlands Institute for Neuroscience, Meibergdreef 47, 1105 BA, Amsterdam, The Netherlands; [127]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; [128]Cluster of Excellence Science of Intelligence, Technische Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany; [129]NeuroMI - Milan Center for Neuroscience, Milan, Italy; [130]Henry H. Wheeler, Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California Berkeley, CA, USA; [131]Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, USA; [132]Scientific and Statistical Computing Core, National Institute of Mental Health, NIH, Bethesda, MD, USA; [133]Department of Psychology, University of Toronto, Canada; [134]Department of Psychology, Florida International University, Miami, FL, USA; [135]Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC, USA; [136]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden; [137]Center for Social and Affective Neuroscience, Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden; [138]Department of Economics and School of Neuroscience, Virginia Tech, Blacksburg, VA USA; [139]Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Forschungszentrum Jülich, Jülich, Germany; [140]Inria, CEA, Université Paris-Saclay, France; [141]Department of Psychology, Columbia University, New York, NY, USA; [142]Department of Management and Engineering, Linköping University, Linköping, Sweden; [143]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden; [144]Center for Cognitive and Decision Sciences, University of Basel, Basel, Switzerland; [145]Department of Psychology, Stanford University, Stanford, CA, USA; [146]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA; [147]Methodology and Statistics Unit, Institute of Psychology, Leiden University, Leiden, The Netherlands; [148]US Combat Capabilities Development Command Army Research Laboratory, USA; [149]University of California Santa Barbara, CA, USA; [150]University of Pennsylvania, PA, USA; [151]Department of Psychology and Neuroscience, Duke University, NC, USA; [152]Center for Cognitive Neuroscience, Duke University, NC, USA; [153]Invicro, London, UK; [154]Faculty of Medicine, Imperial College London, London, UK; [155]Clinical Psychopharmacology Unit, University College London, London, UK; [156]Centre for Human Psychopharmacology, Swinburne University, Hawthorn, VIC, Australia; [157]Department of Management and Marketing, School of Business, University of Dayton, Dayton, OH, USA; [158]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA; [159]Biomedical Engineering Department, University of Connecticut, Storrs, CT, USA

Data analysis workflows in many scientific domains have become increasingly complex and flexible. To assess the impact of this flexibility on functional magnetic resonance imaging (fMRI) results, the same dataset was independently analyzed by 70 teams, testing nine ex-ante hypotheses[1]. The flexibility of analytic approaches is exemplified by the fact that no two teams chose identical workflows to analyze the data. This flexibility resulted in sizeable variation in hypothesis test results, even for teams whose statistical maps were highly correlated at intermediate stages of their analysis pipeline. Variation in reported results was related to several aspects of analysis methodology. Importantly, a meta-analytic approach that aggregated information across teams yielded significant consensus in activated regions across teams. 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 analytic flexibility can have substantial effects on scientific conclusions, and demonstrate factors possibly related to variability in fMRI. The results emphasize the importance of validating and sharing complex analysis workflows, and demonstrate the need for multiple analyses of the same data. Potential approaches 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 (i.e., "researcher's degrees of freedom"[6,7]). Simulation studies show that variability in analytic choices can have substantial effects on results[8], but its degree and impact in practice has not been clear. Recent work in psychology addressed this through a "many analysts" approach[9], in which the same dataset was analyzed by a large number of groups, uncovering substantial variability in behavioral 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), where analysis workflows are complex and highly variable. Our goal was to assess, with the highest possible ecological validity, the degree and impact of analytic 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 analyzing the same dataset. The dataset included fMRI data from 108 individuals, each performing one of two versions of a task previously used to study decision-making under risk[10]. The two versions were designed to address a debate regarding the impact of gain/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 DOI:10.18112/openneuro.ds001734.v1.0.4.

Seventy teams (69 of whom had prior fMRI publications) were provided with the raw data, and an optional preprocessed data (with fMRIprep[13]). They were asked to analyze the data to test nine ex-ante hypotheses (Extended Data Table 1), each consisting of a description of significant 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 based on a whole-brain corrected analysis (yes / no). In addition, each team submitted a detailed report of the analysis methods they had used alongside unthresholded and thresholded statistical maps supporting each hypothesis test (Extended Data Table 2 and Extended Data Table 3a). In order to perform an ecologically valid study testing sources of variability that contribute to published literature "in the wild", instructions to the team were as minimal as possible. The only instructions were to perform the analysis as they

usually would in their own research and report the binary decision based on 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 (Extended Data Table 1 and Figure 1). Only one 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 reporting a different result than the majority of teams (i.e. the absolute distance from consensus). On average across the 9 hypotheses, 20% of teams reported a result that differs 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 analytic choices have a major effect on reported results.

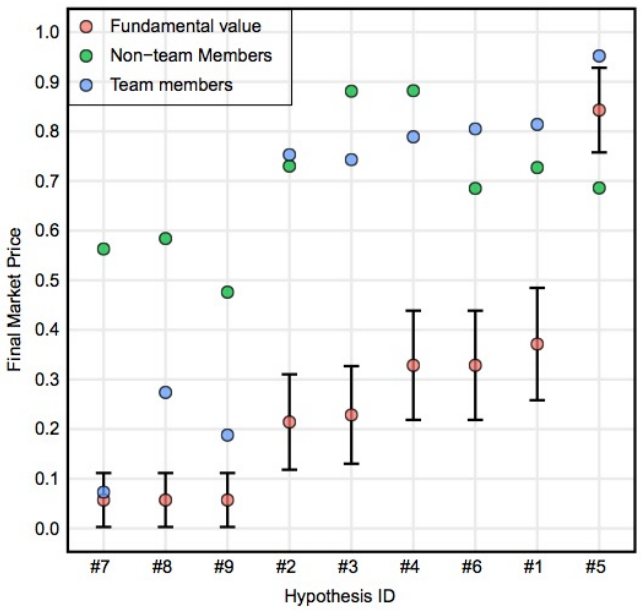


Figure 1: Fraction of teams reporting a significant result and prediction market beliefs. *The figure depicts final market prices for the "team members" (blue dots; N = 83 active traders) and the "non-team members" (green dots; N = 65 active traders) markets as well as the as well as the observed fraction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

Factors related to analytic variability

To examine the sources of the analytic variability in the reported binary results, we analyzed 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 / t) maps of 64 teams were included in the analyses. Fully reproducible code for all analyses of the data reported here are available at DOI: 10.5281/zenodo.3709273.

Variability of reported results. A set of mixed effects logistic regression models identified several analytic 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 FMRIBs Software Library [FSL] smoothest function) was associated with greater likelihood of significant outcomes ($p < 0.001$, delta pseudo-$R^2 = 0.04$; mean FWHM 9.69 mm, range 2.50 - 21.28 mm across teams). Interestingly, while 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 analytic steps beyond explicit smoothing (such as modeling 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], 21 [FSL], 7 [AFNI], 13 [Other]), 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 multiple test correction methods ($p = 0.024$, delta pseudo-$R^2 = 0.02$: N = 48 [parametric], 14 [nonparametric], 2 [other]), with parametric correction methods leading to higher rates of detection than nonparametric

methods. No significant effect was detected for use of standardized preprocessed data versus custom preprocessing pipelines (48% of included teams used fMRIprep; $p = 0.132$) or the modeling 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 analytic variability was further explored by analyzing the submitted statistical maps. The thresholded maps were highly sparse. Binary agreement between thresholded maps over all voxels was relatively high (median percent 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, overlap was very low (median similarity ranging from 0.00 to 0.06 across hypotheses). This may have reflected 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 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 Figure 1).

Variability of unthresholded statistical maps. Analysis of correlation between unthresholded Z-statistic maps across teams demonstrated for each hypothesis a large cluster of teams whose statistical maps were strongly positively correlated with one another (Figure 2 and Extended Data Figure 2). 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 Figure 3a).

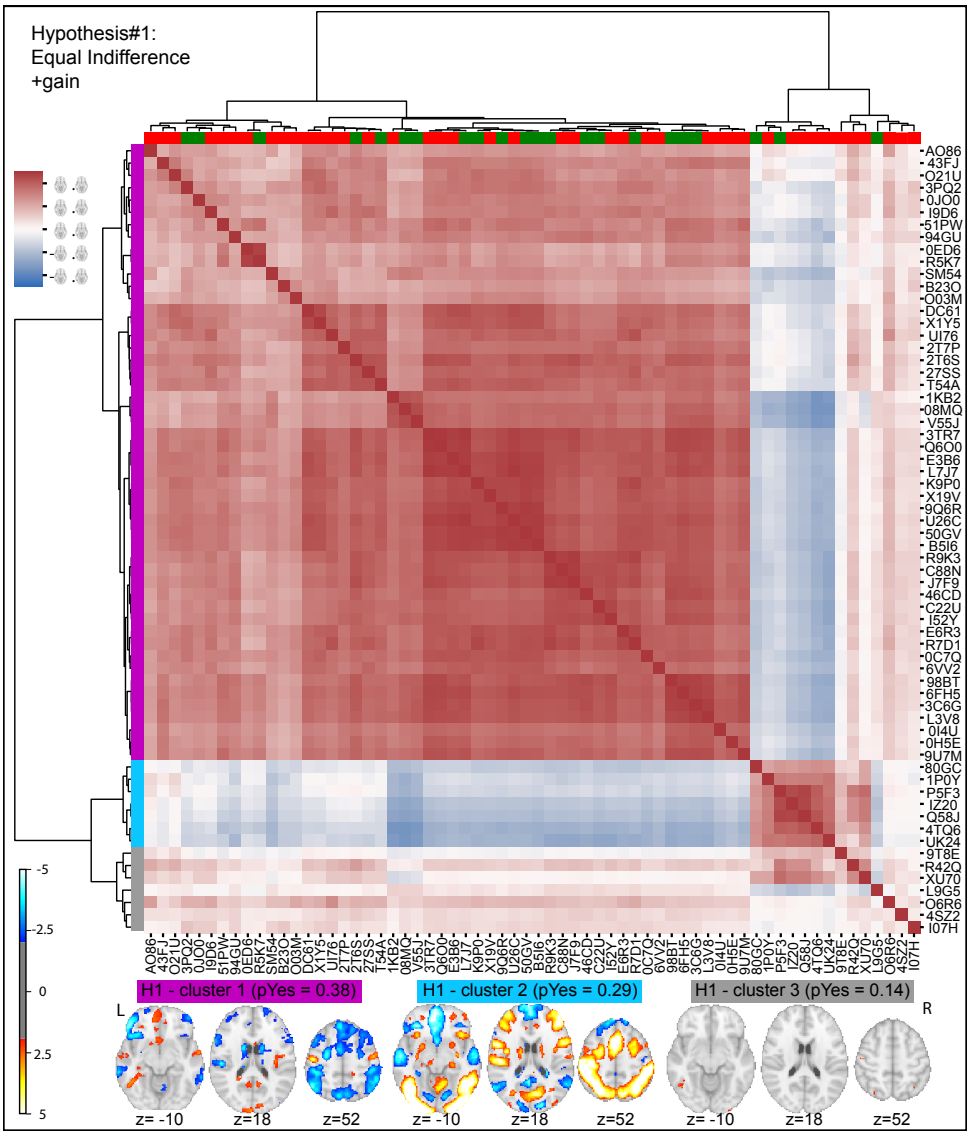


Figure 2. Analytic variability in whole-brain statistical results for Hypothesis 1. Top panel*: Spearman correlation values between whole-brain unthresholded statistical maps for each team (N = 64) were computed and clustered according to their similarity (using Ward clustering on Euclidean distances). Row colors (left) denote cluster membership, while column colors (top) represent hypothesis decisions (green: Yes, red: No). Brackets represent clustering.* ***Bottom panel****: Average statistical maps (thresholded at uncorrected z > 2.0) for each of the three clusters depicted in the left panel. The probability of reporting a positive hypothesis outcome (pYes) is presented for each cluster. Images can be viewed at https://identifiers.org/neurovault.collection:6048.*

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 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; for one additional team, modeling details were not available.

The discrepancy between overall correlations of unthresholded maps and 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 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 Figure 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 #9 after false discovery rate correction (Extended Data Figure 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 all team's results 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, revealing 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 an additional 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 endowed with tokens worth $50 and traded via an online market platform on the fraction of teams reporting a significant result for each hypothesis (i.e. 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 based on their performance in the markets. The analysis of the markets was pre-registered on OSF (https://osf.io/59ksz/). Note that since some analyses were performed on the final market prices (i.e., the markets' predictions), 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 cautiously.

The market's predictions ranged from 0.073 to 0.952 ($m = 0.599$, $sd = 0.325$) in the "team members" market and from 0.476 to 0.882 ($m = 0.690$, $sd = 0.137$) 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 (Figure 1 and Extended Data Table 5a). Spearman correlation between the fundamental values and the markets' predictions 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" prediction market was not driven by over-representation of teams reporting significant results (Supplementary Materials). Predictions in the "team members" market did not significantly differ from those of the "non-team members" (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 their own team's results (Extended Data Table 5b), which may explain why their collective predictions were more accurate than those of non-team members (Figure 1). Additional results are presented in the Supplementary Materials (see also Extended Data Figure 5 and Extended Data Table 5).

Discussion

The analysis of a single functional neuroimaging 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 on a majority of tested hypotheses. For every hypothesis one could find at least four different analysis pipelines used in practice by research groups in the field that 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. Importantly, analyses of the underlying statistical parametric maps on which the

hypothesis tests were based revealed greater consistency than expected from those inferences, with significant consensus in activated regions across groups observed via meta-analysis. Teams with highly correlated underlying unthresholded statistical maps nonetheless reported divergent hypothesis outcomes (Figure 2). Detailed analysis of the workflow descriptions and statistical results submitted by the analysis teams identified several common analytic 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), choices of analysis software and 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 leading to statistical maps that were anticorrelated with the majority. Prediction markets that were performed on the outcomes of analyses demonstrated that researchers generally overestimated the likelihood of significant results across hypotheses, even by those researchers who had analyzed the data themselves, reflecting substantial optimism bias by researchers in the field.

The substantial amount of analytic variability, leading to 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. In addition, publicly sharing data and analysis code should become common practice, to enable others to run their own analysis with the same data or validate the code used. These practices alongside the use of pre-registration[16] or registered reports[17] will reduce researchers' degrees of freedom but would not prevent analytic variability as

demonstrated here; however, they would ensure that the impact 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 analyzed using multiple analysis pipelines, preferably by more than one research team. Achieving such "multiverse analysis" at scale will require the development of automated statistical analysis tools (e.g.[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 in order 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 analytic results. Nonetheless, we maintain that fMRI can provide reliable answers to scientific questions, as strongly demonstrated in the meta-analytic 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 functional neuroimaging dataset, it seems highly likely that similar variability will be present for other fields of research where 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. Importantly, transparent community-wide self-assessment scientific projects, such as the current one, are definitive evidence of the researchers' awareness of reproducibility concerns and desire to assess their impact and improve practices accordingly (for additional discussion see Supplementary Discussion).

References

1. Botvinik-Nezer, R. et al. fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study. Scientific Data 6, 1–9 (2019).
2. Dreber, A. et al. Using prediction markets to estimate the reproducibility of scientific research. Proceedings of the National Academy of Sciences 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. Nature Human Behaviour 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, 1–12 (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, 1–13 (2012).
9. Silberzahn, R. et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science 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. Proceedings of the National Academy of Sciences 107, 3788–3792 (2010).
12. Canessa, N. et al. The Functional and Structural Neural Basis of Individual Differences in Loss Aversion. Journal of Neuroscience 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, 1–9 (2015).
16. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. Proceedings of the National Academy of Sciences 2017, 201708274 (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. De La Vega Tal Yarkoni Russell Poldrack Krzysztof Gorgolewski, C. M. A. FitLins: Reproducible model estimation for fMRI. in.
19. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. (2015) doi:10.2139/ssrn.2694998.

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).

Methods

fMRI dataset

In order 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 pre-defined hypotheses. The full dataset is publicly available in the Brain Imaging Data Structure (BIDS)[23] on OpenNeuro (DOI: 10.18112/openneuro.ds001734.v1.0.4) and is described in detail in a Data Descriptor[1].

Shortly, the fMRI dataset consisted of data from 108 participants performing 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, where people tend to be more sensitive to losses compared to equal-sized gains[24]); 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 version 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 Scientific 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 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. We have complied with all relevant ethical regulations.

Pre-defined hypotheses

Previous studies with the mixed gambles task suggested that activity in the vmPFC 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 (i.e., equal indifference versus equal range). Each analysis team tested nine pre-defined 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 / loss amount) and a specific group (equal indifference / equal range, or a comparison between the two groups). Therefore, for each hypothesis, the maximal 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 based on a whole-brain analysis (and not on a region of interest based analysis, as sometimes used in fMRI studies).

Analysis teams recruitment and instructions

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 analyze 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 based on a whole-brain analysis. While 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 regions of interest. Each team also filled in a full report of the analysis methods used (following COBIDAS guidelines[25]) and created a collection on NeuroVault[15] [RRID:SCR_003806] with one unthresholded and one thresholded statistical maps for each hypothesis, on which their decisions were based (teams could optionally include additional maps in their collection; see Extended Data Table 3a for collections' links). For each result (i.e., 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 Table 1 and Extended Data Table 2. In order to measure 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 based on their own criteria.

Seventy of the 82 teams submitted their results and reports by the final deadline (March 15th, 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. In order to avoid identification of the teams, each team was provided with a unique random 4-character team ID.

Overall, 180 participants were part of NARPS analysis teams. Out of 70 analysis teams, five teams consisted of one member, 20 teams consisted of two members and 45 teams consisted of three members. Out of the 180 team members, there were 62 principal investigators (PIs), 43 post-doctoral fellows, 53 graduate students and 22 members from other positions (e.g. data scientists or research analysts).

Factors related to analytic variability

In order 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 v2.0.3 (DOI: 10.5281/zenodo.3709273). We performed exploratory analyses of the relation between the reported hypothesis outcomes and several analytic 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 FSL's smoothest function), use of standardized preprocessing, software package, method of correction for multiple comparisons and modeling of head movement. The teams were modeled 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 modeling, 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 based on 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 explore 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, 91x109x91, 2mm isotropic) using nilearn[28] [RRID:SCR_001362]. For unthresholded maps, we used 3rd order spline interpolation; for thresholded maps, we used linear interpolation and then thresholded at 0.5, to prevent artifacts that appeared when using nearest neighbor interpolation. Of the 69 teams included in the analyses, unthresholded maps of five teams and thresholded maps of four teams were excluded from the

image-based analyses (see Extended Data Table 3b for details). Since 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 percent agreement, i.e. the percent 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 nonzero 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 (ALE; see Supplementary Materials)[30,31].

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 based on visual examination of the dendrograms. A separate mean statistical map was then created for the teams in each cluster (see Figure 2 and Extended Data Figure 2). Drivers of map similarity were further assessed by modeling the median correlation distance of each team from the average pattern as a function of several analysis decisions (e.g. smoothing, whether or not the data preprocessed with fMRIprep were used, etc.).

To assess the impact of variability in thresholding methods and anatomical definitions across teams, unthresholded Z maps for each team were thresholded using a common approach. 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 false discovery rate 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 region of interest for each hypothesis. The regions of interest for the ventral striatum and amygdala were defined anatomically based on the Harvard-Oxford anatomical atlas. Since 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-analytic map obtained from neurosynth.org[33] for the search term "ventromedial prefrontal".

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

Prediction markets

The second main goal of the Neuroimaging Analysis Replication and Prediction Study (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 $50) and traded on the outcome of the hypotheses via 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, i.e., traders were paid based on their performance in the markets.

Recruitment. For the "non-team members" prediction market, we invited participants via social media (mainly Facebook and Twitter) and e-mails. The invitation contained a link to an online form on the NARPS website (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, via email directing 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 initially were not aware that they would be invited to participate in a separate prediction market after they had analyzed the data. The decision to implement a second market, consisting of traders with partial information about the fundamental values (i.e., 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 / no). The invitations to participate in the prediction markets were first distributed on April 9, 2019; the registration closed on April 29, at 4pm UTC. Upon closure of the registration, all participants received a personalized email containing a link to the web-based market software and their login-credentials. The prediction markets opened on May 2, 2019 at 4pm UTC and closed on May 12, 2019 at 4pm 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 (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 (i.e., 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 (i.e., 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 May 6. The error was corrected immediately on the same day and all participants were informed about the mistake on our part via a personal email notification (on May 6, 2019, 3:28pm 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 are 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 (i.e., hypotheses). Unlike prediction markets on binary outcomes (e.g., 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] (though see[37] and[38] for caveats), 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 based on 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) 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 Camerer et al.[3].

Incentivization. Once the markets had been closed, the true "fundamental value" (FV) for each asset (i.e., 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 (i.e., 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 (i.e., 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 USD at a rate of 1 Token = \$0.5. The final payments were transferred to participants during the months May to September 2019 in form of Amazon.com giftcards, Amazon.de giftcards, or PayPal payments, depending on the preferred mode of payment indicated by the participants upon registration for participating in 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 (i.e., hypotheses) ranged from 46 to 76 (m = 56.4, sd = 8.9) in the "team members" set of markets and from 35 to 58 (m = 47.1, sd = 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-docs 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. The majority of the participants resided in Europe (46.3%) and North America (46.3%).

Pre-Registration. All analyses of the prediction markets data reported were pre-registered at https://osf.io/pqeb6/. The pre-registration was completed after the markets opened, but before the markets closed. Only one member of the NARPS research group, Felix 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 pre-registration. Only two members of the NARPS research group, Rotem Botvinik-Nezer and Tom Schonberg, had any information about the results reported by the 70 analyses teams before the prediction markets closed. Neither of them were involved in writing the pre-registration either.

For additional details on the prediction markets, see the Supplementary Materials.

Data and code availability

The full fMRI dataset is publicly available on OpenNeuro (DOI: 10.18112/openneuro.ds001734.v1.0.4) and is described in details in a Data Descriptor[1].

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://cloud.docker.com/repository/docker/poldrack/narps-analysis - tag:paper_analysis). 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. Imaging analysis code was independently reviewed by an expert who was not involved in writing the original code. Prediction market analyses were performed

using R v3.6.1; packages were installed using the checkpoint package, which reproducibly installs all package versions as of a specified date (August 13th, 2019). Analyses reported in this manuscript were performed using code release v2.0.3 (DOI: 10.5281/zenodo.3709273).

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 2a.

Interested 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 imaging analyses are available for anonymous download at DOI:10.5281/zenodo.3709275.

Although not required to, several analysis teams also publicly shared their analysis code. Extended Data Table 3d includes these teams along with the link to their code.

Methods References

23. Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 3, 1–9 (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. 8, 1–10 (2014).
29. Hughett, P. Accurate Computation of the F-to-z and t-to-z Transforms for Large Arguments. J. Stat. Softw. 23, (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. Proceedings of the National Academy of Sciences 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. (2006) doi:10.3386/w12200.
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. The Journal of Prediction Markets 1, 3–15 (2007).
40. Chen, Y. Markets as an information aggregation mechanism for decision support. (Penn State University, 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; granted to Tom Schonberg) and the Swedish Foundation for Humanities and Social Sciences (NHS14-1719:1). Hosting of the data on OpenNeuro supported by NIH grant R24MH117179. Thanks to Michael C. Frank, Yaniv Assaf and Nathaniel Daw for helpful comments on an earlier draft. Thanks to the Texas Advanced Computing Center for providing computing resources for preprocessing of the data, and the Stanford Research Computing Facility for hosting the data. Thanks to Dana Roll for assisting with data processing. A. Dreber thanks the Knut and Alice Wallenberg Foundation and the Marcus and Marianne Wallenberg Foundation (A. Dreber is a Wallenberg Scholar), the Austrian Science Fund (FWF, SFB F63) and the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser). D. Wisniewski was supported by the Research

Foundation Flanders (FWO, fwo.be), 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 acknowledges the University of Basel Research Fund for Junior Researchers. C.B. Calderon was supported by grant 12O7719N from the Research Foundation Flanders. E. Lesage 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 HPC usage. C. Lamm was supported by the Viennese Science and Technology Fund (WWTF VRG13-007) and Austrian Science Fund (FWF P 32686). A. Losecaat Vermeer was supported by the Viennese Science and Technology Fund (WWTF VRG13-007). L. Zhang was supported by 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. Pischedda 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).

Author contributions

- NARPS management team: R. Botvinik-Nezer, F. Holzmeister, C.F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R.A. Poldrack and T. Schonberg.
- fMRI dataset- experiment design: R. Iwanir, J. Durnez, R.A. Poldrack and T. Schonberg.
- fMRI dataset- data collection: R. Iwanir and T. Schonberg.
- fMRI dataset- preprocessing, quality assurance and data sharing: R. Botvinik-Nezer, K. Gorgolewski, R.A. Poldrack and T. Schonberg.
- Analysis teams- recruitment, point of contact and management: R. Botvinik-Nezer, R.A. Poldrack and T. Schonberg.

- Analysis teams- analysis of the submitted results and statistical maps: R.A. Poldrack, T.E. Nichols, J.A. Mumford, J-.B. Poline, A. Perez, R. Botvinik-Nezer, and T. Schonberg.
- Code review: T. Glatard. and K. Dadi.
- Prediction markets- design and management: F. Holzmeister, C.F. Camerer, A. Dreber, J. Huber, M. Johannesson and M. Kirchler
- Prediction markets- analysis: F. Holzmeister, R. Botvinik-Nezer, C.F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, S. Kupek, R.A. Poldrack and T. Schonberg.
- Writing the manuscript: R. Botvinik-Nezer, F. Holzmeister, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, T.E. Nichols, R.A. Poldrack and T. Schonberg.
- Participated as members of analysis teams and reviewed and edited the manuscript: R.A. Adcock, P. Avesani, B.M. Baczkowski, A. Bajracharya, L. Bakst, S. Ball, M. Barilari, N. Bault, D. Beaton, J. Beitner, R.G. Benoit, R.M.W.J. Berkers, J.P. Bhanji, B.B. Biswal, S. Bobadilla-Suarez, T. Bortolini, K.L. Bottenhorn, A. Bowring, S. Braem, H.R. Brooks, E.G. Brudner, C.B. Calderon, J.A. Camilleri, J.J. Castrellon, L. Cecchetti, E.C. Cieslik, Z.J. Cole, O. Collignon, R.W. Cox, W.A. Cunningham, S. Czoschke, K. Dadi, C.P. Davis, A. De Luca, M.R. Delgado, L. Demetriou, J.B. Dennison, X. Di, E.W. Dickie, E. Dobryakova, C.L. Donnat, J. Dukart, N.W. Duncan, A. Eed, S.B. Eickhoff, A. Erhart, L. Fontanesi, G.M. Fricke, S. Fu, A. Galván, R.i Gau, S. Genon, E. Glerean, J.J. Goeman, S.A.E. Golowin, C. González-García, K.J. Gorgolewski, C.L. Grady, M.A. Green, J.F. Guassi Moreira, O. Guest, S. Hakimi, J.P. Hamilton, R. Hancock, G. Handjaras, B.B. Harry, C. Hawco, P. Herholz, G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P. Hu, S.A. Huettel, M.E. Hughes, V. Iacovella, A.D. Iordan, P.M. Isager, A.I. Isik, A. Jahn, M.R. Johnson, T. Johnstone, M.J.E. Joseph, A.C. Juliano, J.W. Kable, M. Kassinopoulos, C. Koba, X.-Z. Kong, T.R. Koscik, N.E. Kucukboyaci, B.A. Kuhl, A.R. Laird, C. Lamm, R. Langner, N. Lauharatanahirun, H. Lee, S. Lee, A. Leemans, A. Leo, E. Lesage, F. Li, M.Y.C. Li, P. Cheng Lim, E.N. Lintz, S.W. Liphardt, A.B. Losecaat Vermeer, B.C. Love, M.L. Mack, N. Malpica, T. Marins, C. Maumet, K. McDonald, J.T. McGuire, H. Melero, A.S. Méndez Leal, B. Meyer, K.N. Meyer, G. Mihai, G.D. Mitsis, J. Moll, D.M. Nielson, G. Nilsonne, M.P. Notter, E. Olivetti, A.I. Onicas, P. Papale, K.R. Patil, J.E. Peelle, D. Pischedda, Y. Prystauka, S. Ray, P.A. Reuter-Lorenz, R. C Reynolds, E. Ricciardi, J.R. Rieck, A.M. Rodriguez-Thompson , A. Romyn, T. Salo, G.R. Samanez-Larkin, E. Sanz-Morales, M.L. Schlichting, D.H. Schultz, Q. Shen, M.A. Sheridan, J.A. Silvers, K. Skagerlund, A. Smith, D.V. Smith, P. Sokol-Hessner, S.R. Steinkamp, S.M. Tashjian, B. Thirion, J.N. Thorp, G. Tinghög, L. Tisdall, S.H. Tompson, C. Toro-Serey, J.J. Torre Tresols, L. Tozzi, V. Truong, L. Turella, A.E. van 't Veer, T. Verguts, J.M. Vettel, S. Vijayarajah, K. Vo, M.B. Wall, W.D. Weeda, S. Weis, D.J. White, D. Wisniewski, A. Xifra-Porxas, E.A. Yearling, S. Yoon, R. Yuan, K.S.L. Yuen, L. Zhang, X. Zhang, J.E. Zosky

Competing interest statement

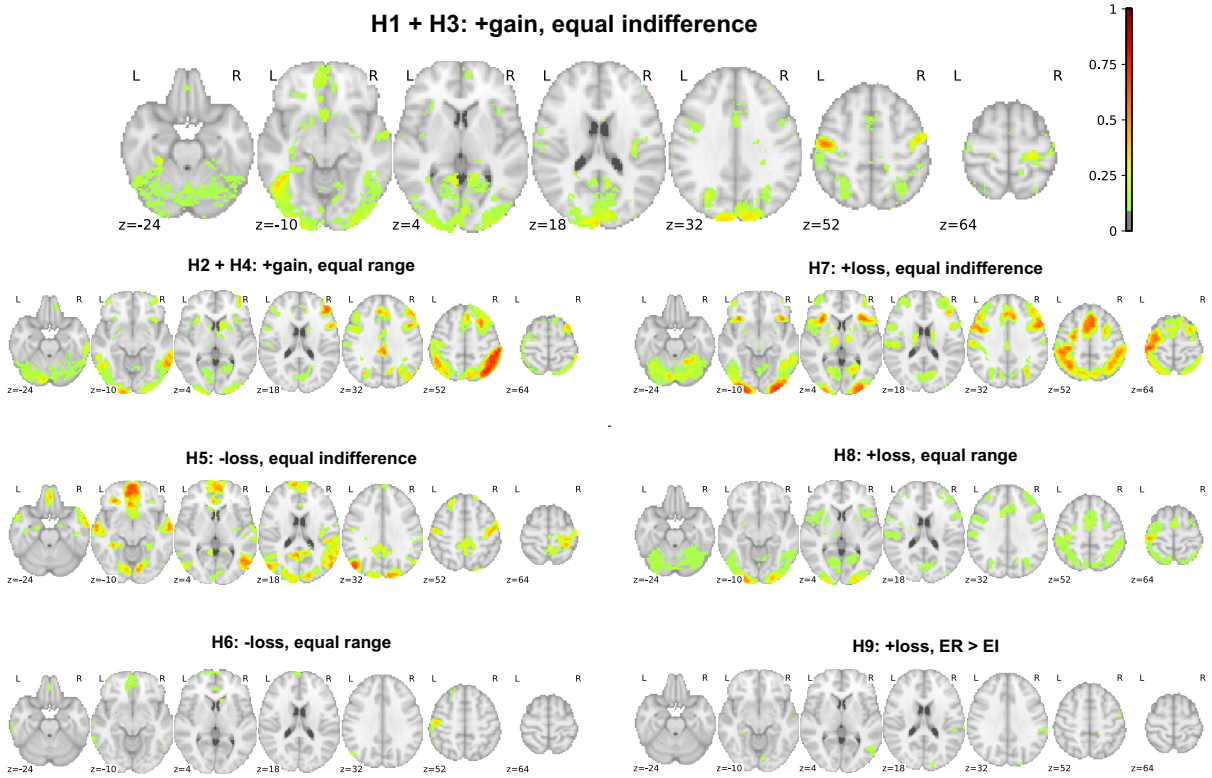
The authors declare no competing interests.

Additional information

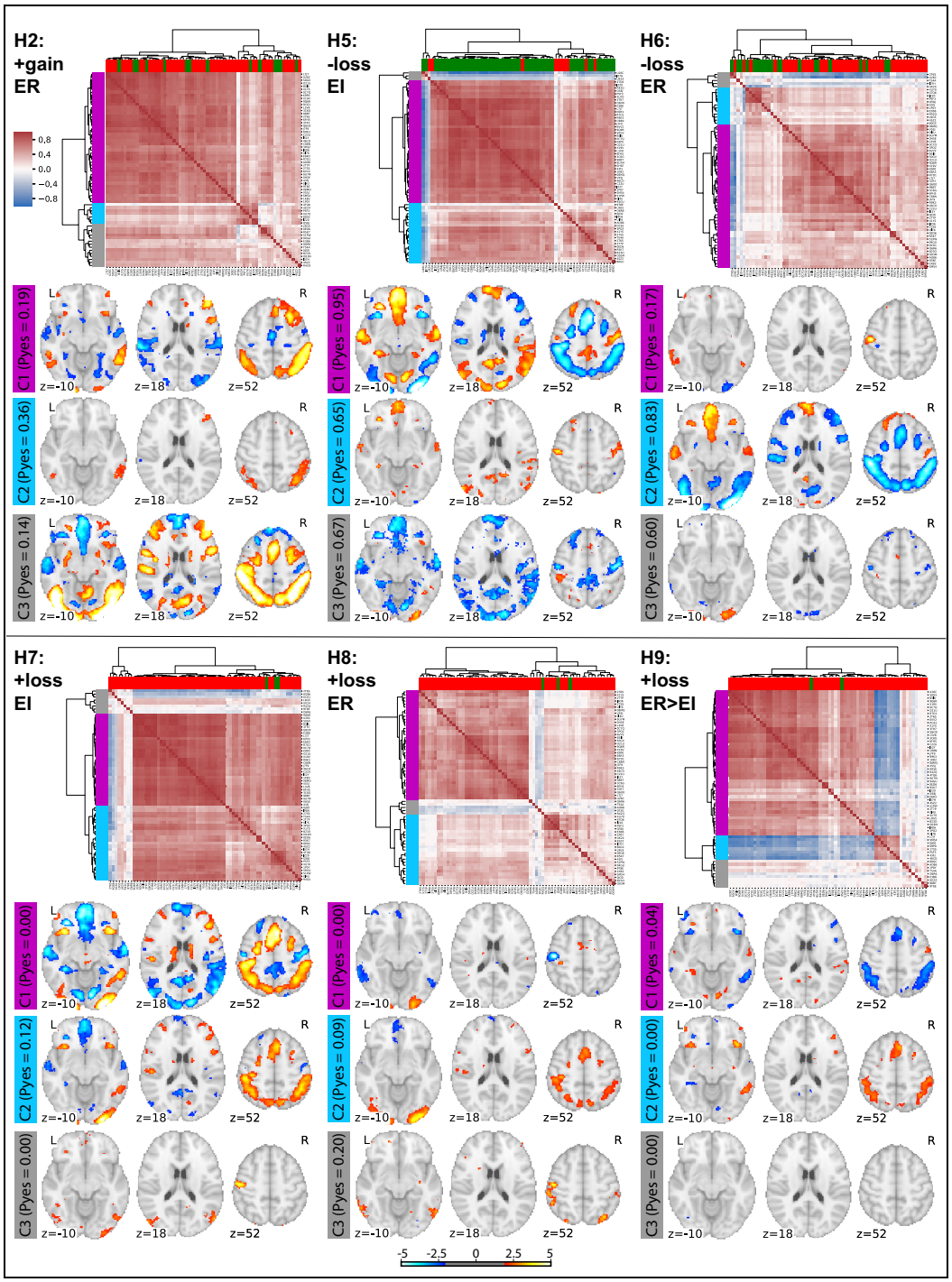
Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Tom Schonberg or Russell A. Poldrack.

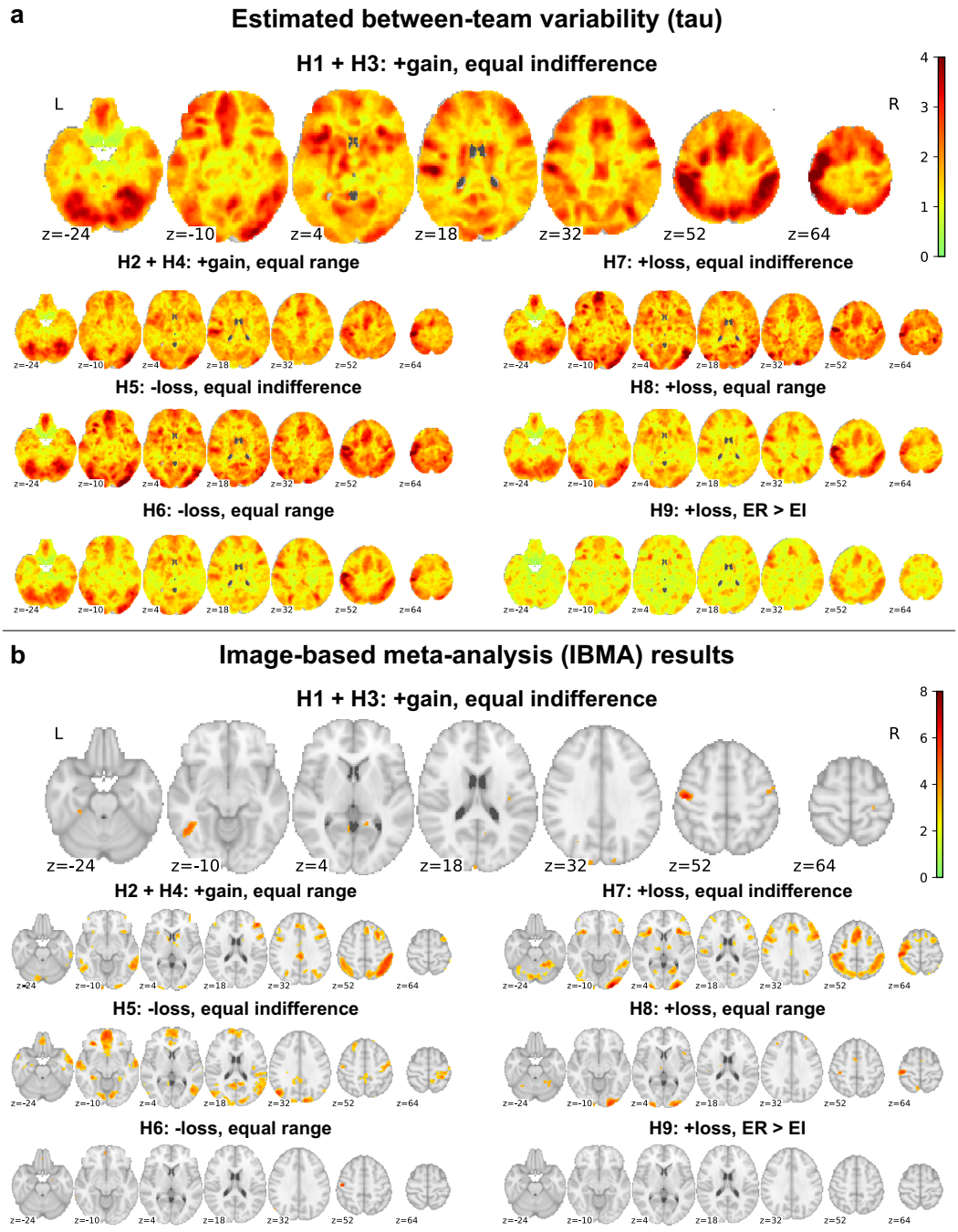
Extended Data Figures



Extended Data Figure 1 | Voxels overlap. Maps showing at each voxel the proportion of teams (out of N = 65 teams) reporting significant activations in their thresholded statistical map, for each hypothesis (labeled H1 - H9), thresholded at 10% (i.e., voxels with no color were significant in fewer than 10% of teams). +/- refers to direction of effect, gain/loss refers to the effect being tested, and equal indifference (EI) / 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 the hypotheses are for the same contrast and experimental group, but for different regions (see Extended Data Table 1). Images can be viewed at https://identifiers.org/neurovault.collection:6047

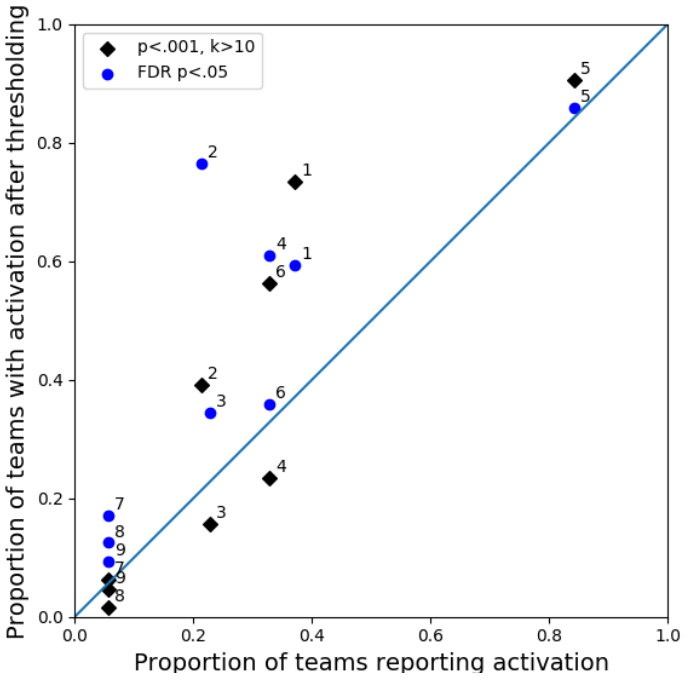


Extended Data Figure 2 | Variability of whole-brain unthresholded maps for hypotheses 2, 4 - 9. For each hypothesis, we present a heatmap based on Spearman correlations between unthresholded statistical maps (N = 64), clustered according to their similarity, and the average of unthresholded images for each cluster (cluster colors in titles refer to colors in left margin of heatmap). Green / red at the columns represent binary results (significant / not significant, respectively) reported by the analysis teams; row colors represent cluster membership. Maps are thresholded at an uncorrected value of $Z >$ 2 for visualization. Unthresholded maps for Hypothesis #2 and Hypothesis #4 are identical (as they both relate to the same contrast and group, but different regions), and the colors represent reported results for Hypothesis #2. For Hypotheses #1 and #3 see Figure 2.



Extended Data Figure 3 | Variability and consensus of unthresholded statistical maps (N = 64). (**a**) Maps of estimated between-team variability (tau) at each voxel for each hypothesis. Images can be viewed at https://identifiers.org/neurovault.collection:6050. (**b**) Image-based meta-analysis (IBMA) results. 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 due to the same underlying data (see Methods). Maps are presented for each hypothesis showing voxels (in color) where the group statistic was significantly greater than zero after voxelwise correction for false discovery rate ($p < 0.05$). Color bar reflects statistical value (Z) for the meta-analysis. Images can be viewed at https://identifiers.org/neurovault.collection:6051.
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 for different regions (see Extended Data Table 1).

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	0.359a	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

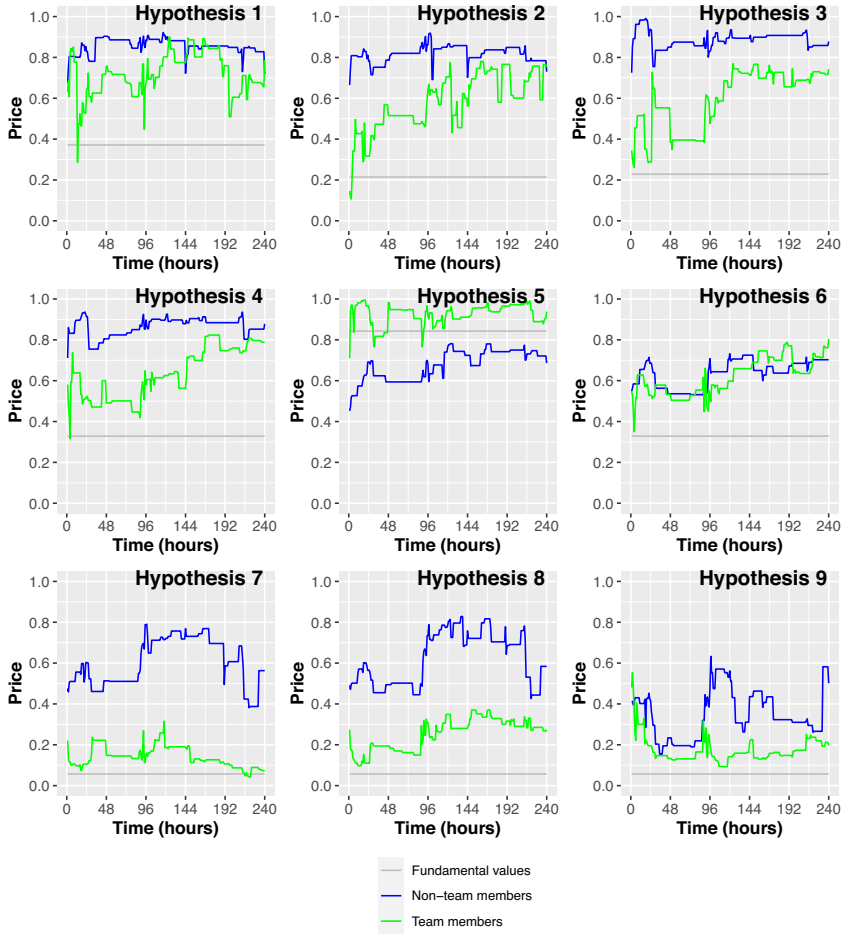
Extended Data Figure 4 | Results of the consistent thresholding and ROI selection analysis ($N = 64$). (**a**) Activation for each hypothesis as determined using consistent thresholding (black: $p < 0.001$ and *cluster size* > 10 voxels; blue: FDR correction with $p < 0.05$) and ROI selection across teams (y-axis), versus 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 uncorrected ($p <$ 0.001, cluster size $k > 10$) and false discovery rate correction ($pFDR < 5\%$) and common anatomical regions of interest 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 also presented, thresholded at pFDR < 5% as well.

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 Figure 5 | Prediction markets over time (*N* = 240 observations [10 days X 24 hours]). (**a**). Panel regressions. The table summarizes the results of pre-registered fixed-effects panel regressions of the predictions absolute errors (i.e., 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 are 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 prediction market prices per hour separated for the two prediction markets for the time the markets were open (10 days, i.e., 240 hours). The gray line indicates the actual share of analysis teams reporting a significant result for the hypothesis (i.e., the fundamental value).

Extended Data tables

	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)

Extended Data Table 1 | Hypotheses and results
Each hypothesis is described along with the fraction of teams reporting a whole-brain corrected significant result (out of $N = 70$ teams) and two measures reported by the analysis teams for the specific hypothesis (both rated 1-10): (1) How confident are you about this result? (2) How similar do you think your result is to the other analysis teams? For these ordinal measures, median values are presented along with the median absolute deviation in brackets. See Supplementary Materials for analysis of the confidence level and similarity estimation.

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

Extended Data Table 2 | Results submitted by analysis teams*

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] to 10 [extremely]) 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).

* Three teams changed their decisions after the end of the project. Team L3V8 changed their decision regarding Hypothesis #6 from "yes" to "no". Team VG39 changed their decisions regarding Hypotheses #3, #4 and #5 from "yes" to "no". Team U26C changed their decision regarding Hypothesis #5 from "yes" to "no". Results along the paper and in this table reflect the final results as they were reported at the end of the project (i.e., before this change), as prediction markets were based on those results.

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

Extended Data Table 3 | Data links and analysis related tables

(**a**) Numbers of public NeuroVault collections of all analysis teams (full link: https://neurovault.org/collections/<*insert collection number here*>/). (**b**) Description of teams excluded from the analyses of statistical maps. (**c**) Summary of mixed-effects logistic regression modeling of decision outcomes (N = 64 per hypothesis) as a function of different factors including the hypothesis (1-9) and various aspects of statistical modeling (for modeling details see https://github.com/poldrack/narps/blob/master/ImageAnalyses/DecisionAnalysis.Rmd). (**d**) Links to shared analysis code of some of the analysis 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

Extended Data Table 4 | Variability of statistical maps across teams
(**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).

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

Extended Data Table 5 | Prediction markets results and additional data

(**a**). A summary of the prediction market results. FV indicates the fundamental value, i.e., the actual fraction of teams (out of $N = 70$ teams) reporting significant results for the hypothesis. 95% 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 section of the table reports 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 lower section reports 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. Average holdings if (in)consistent refer to the mean final holdings, separated for consistent and inconsistent traders. (**c**) The table depicts additional data for each of the nine hypotheses. Tokens invested indicates the average number of tokens invested per transaction and Volume (Shares) refers to the mean number of shares bought or sold per transaction. Transactions describes the overall number of transactions recorded and # Traders refers to the number of traders who bought or sold shares of the particular asset at least once.