A Supplement to methods

A.1 HCP experimental task details

Working Memory Each of the two runs of the working memory task consisted of eight task (25 s each) and four fixation blocks (15 s each). In each of the task blocks, particpants saw images of one of four different stimulus types (namely, images of body parts, faces, places or tools). These four stimulus types are known to reliably engage distinct cortical regions (Downing et al., 2001) across subjects Peelen and Downing (2005) and time (Fox et al., 2009). Half of the task blocks used a 2-back working memory task (participants were asked to respond "target" when the current stimulus was the same as the stimulus 2 back) and the other half a 0-back working memory task (a target stimulus was presented at the beginning of each block and participants were asked to respond "target" whenever the target stimulus was presented in the block). Each task block consisted of 10 trials (2.5 s each). In each trial, a stimulus was presented for 2 s followed by a 500 ms interstimulus interval (ISI). We were not interested in identifying any effect of the N-back task condition on the evoked brain activity and therefore pooled the data of both N-back conditions.

Gambling Participants played a card guessing game in which they were asked to guess the number on a mistery card. The potential card numbers ranged from 1 to 9 and participants were asked to indicate whether they think that the number is going to be above or below 5. Participants received feedback in form of a number on the card. Importantly, the number on the card was dependent on whether the respective trial was marked as a reward, loss, or neutral trial. In addition to the number, the feedback included a green arrow pointing upwards with "1" for reward trials or a red arrow pointing downwards next to "−0.50" for loss trials or the number "5" and a gray double headed arrow for neutral trials. Participants had 1.5 s to indicate a guess (during this time a "?" was presented), while the subsequent feedback was presented for 1.0 s. In addition, there was a 1.0 s intertrial interval with a "+" on the screen. The task was presented in blocks that each included eight trials that were either mostly reward (6 rewards trials that were pseudo-randomly interleaved with either 1 neutral and 1 loss trial, 2 neutral trials, or 2 loss trials) or mostly loss (6 loss trials interleaved with either 1 neutral and 1 reward trial, 2 neutral trials, or 2 reward trials) trials. In each of the two fMRI runs there were 2 mostly reward and 2 mostly loss blocks, interleaved with 4 fixation blocks (15 s each, during which a "+" is presented on the screen). All participants were provided with money as a result of completing the experiment. The amount they received was standardized due to the fixed nature of the experiment.

Motor Participants were presented with visual cues that asked them to tap their left or right fingers, squeeze their left or right toes, or move their tongue. The task was presented in blocks of 12 s that each included only one movement type (10 movements). Each block was preceded by a 3 s cue. In each of the two fMRI runs, 13 blocks were presented with 2 blocks for tongue movements, 4 blocks for hand movements (2 left, 2 right), and 4 blocks for foot movements (again, 2 left and 2 right). In addition, three 15 s fixation blocks were included in each run.

Language This task consisted of two runs that each interleaved 4 blocks of a story task and 4 blocks of a math task. In the story task, participants were presented with brief auditory stories (5-9 sentences) that were adapted from Aesop's fables. After each story, a 2-alternative forced-choice question asked the particpant about the topic of the story. In the math task, particpants were similarly presented with an auditory math problem that asked them to complete 2-alternative forced choice addition or subtraction problems. For example, participants heard the operation "fourteen plus twelve", followed by "equals" and then two choice alternatives ("twenty-nine or twenty-six"). Participants indicated with a button press whether they choose the first or second answer. The lengths of the blocks varied (with an average of approximately 30 s per block), but

the task was designed in such a way that the math task blocks matched the length of the story task blocks (with some additional math trials at the end of a block if needed to complete the 3.8min run).

Social Participants were presented with video clips (20 s each) that showed objects (squares, circles, triangles) that either interacted in some way or were moving randomly. After each video clip, participants indicated whether they think that the objects had a social interaction (an interaction that appears as if the objects are taking into account each other's feelings or thoughts), they are not sure, or they think the objects did not interact. Each of the two fMRI runs included 5 video blocks (2 with interaction and 3 without in one run and 3 with interaction and 2 without in the other run) as well as 5 15 s fixation blocks.

Relational In this task, participants saw stimuli that were composed of six different shapes that were filled with one of six different textures. In the relational task condition, 2 pairs of objects were presented, one at the top of the screen and the other at the bottom. Participants were told that they should first decide what dimension (shape or texture) differs across the top pair of objects and then whether the bottom pair of objects differs along the same dimension. In the matching condition, participants were shown two objects at the top of the screen and one at the bottom. A word in the middle of the screen then indicated whether participants should decide if the bottom object matched either of the two top objects on the "shape" or "texture" dimension. In the relational condition, stimuli were presented for 3500 ms, with a 500 ms intertrial interval and four trials per block. In the matching condition, stimuli were presented for 2800 ms, with a 400 ms intertrial interval, and a total of five trials per block. Each block lasted a total of 18 s. In each of the two fMRI runs three relational blocks, three matching blocks and three fixation blocks (16 s each) were presented.

Emotion In emotion trials, participants were presented with with two faces at the bottom of the screen and one face at the top. These faces had an either angry or fearful expression. The participants were asked to decide which of the two faces on the bottom matches the face at the top. In neutral trials, participants were asked to decide which of two shapes at the bottom of the screen matches a shape that is presented at the top. In this task, trials were presented in blocks of six trials of the same task (face or shape). In each trial, the stimulus was presented for 2 s in addition to a 1 s intertrial interval. Each block was further preceded by a 3 s cue for the task (shape or face). Each of the two fMRI runs included three face and three shape blocks. Due to a bug in the experiment script, the experiment stopped before the final three trials of the last block of each trial (for further details on this bug, see Barch et al. (2013)).

A.2 GLM analysis details

FMRI Our GLM subject-level analyses of the fMRI data included one predictor for each of the four cognitive states in the design matrix (each representing a box-car function for the occurrence of a cognitive state). We convolved these predictors with a canonical glover haemodynamic response function (HRF; Lindquist et al., 2009) as implemented in Nilearn 0.8.0 (Abraham et al., 2014), to generate the model predictors. We added temporal derivative terms derived from each predictor, an intercept and an indicator of the experiment run to the design matrix, which we all treated as confounds of no interest. The derivative terms were computed by the use of the cosine drift model as implemented in Nilearn 0.8.0 (Abraham et al., 2014).

To generate a set of group-level brain maps with the GLM, we computed a second-level GLM contrast by the use of the standard two-stage procedure for a random-effects group-level analysis, as proposed by Holmes and Friston (1998). Here, the subject-level regression coefficients β are

treated as random effects in a second-level linear contrast analysis, where the distribution of first-level β-contrasts is assessed. Contrasts were computed between each cognitive state and all others. The resulting group-level brain maps show the z-scores resulting from this test.

Relevances Our GLM analyses of the relevance data resulting from the application of the LRP technique to DeepLight's decoding decisions (for an overview of the LRP technique, see section 4.4 of the main text) included one predictor for each of the four cognitive states in the data (each representing a box-car function for the occurrence of a cognitive state). Our previous analyses have indicated that DeepLight's relevance data show a similar temporal evolution as the HRF (see Fig. 6 of Thomas et al., 2019a). For this reason, we next convolved the predictors with a canonical glover HRF (Lindquist et al., 2009), as implemented in Nilearn 0.8.0 (Abraham et al., 2014), to generate a set of model predictors.

We further added temporal derivative terms derived from each predictor, an intercept and an indicator of the experiment run to the design matrix. The temporal derivative terms were computed by the use of the cosine drift model as implemented in Nilearn 0.8.0 (Abraham et al., 2014). Additionally, we added one regressor to the design matrix indicating the total sum of relevance values contained in each fMRI volume (i.e., TR), to account for the variability in the sum of relevance values between TRs resulting from variability in the certainty of DeepLight's predictions (for an overview of the LRP technique, see section 4.4 of the main text). To also account for non-linear relationships between this regressor and the relevance values, we added regressors for the first derivative of the relevance sums, the squared relevance sums, and the first derivative of the squared relevance sums to the design matrix. All of these predictors were treated as confounds of no interest.

Lastly, we added two regressors to the design matrix indicating whether DeepLight correctly or incorrectly identified the cognitive state of each TR (again in form of two box-car functions). Importantly, we included these two predictors in each computed contrast, by contrasting each cognitive state against all other states and by contrasting correct versus incorrect predictions (e.g., to compute a contrast for the body state in the HCP-WM task (see section 4.1.1 of the main text), we would set the contrast vector to: 3, -1, -1, -1, 1, -1 for the predictors: body, face, place, tool, correct, incorrect).

To generate a set of group-level brain maps with the GLM, we computed a second-level GLM contrast by the use of the standard two-stage procedure for a random-effects group-level analysis, as proposed by Holmes and Friston (1998). Here, the subject-level regression coefficients β are treated as random effects in a second-level linear contrast analysis, where the distribution of first-level β-contrasts is assessed. The resulting group-level brain maps show the Z-values resulting from this test.

A.3 FMRIPrep details for Multi-task data

This dataset was processed using *fMRIPrep* 20.0.5 (Esteban et al. (2019); Esteban et al. (2018); RRID:SCR_016216), which is based on *Nipype* 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). Volume-based spatial normalization to two standard

spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym], *ICBM 152 Nonlinear Asymmetrical template version 2009c* [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

Functional data preprocessing For each of the 18 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using `flirt` (FSL 5.0.9, Jenkinson and Smith, 2001) with the boundary-based registration (Greve and Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). BOLD runs were slice-time corrected using `3dTshift` from AFNI 20160207 (Cox and Hyde, 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin6Asym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded

(volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

The above boilerplate text was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts *unchanged*. It is released under the CC0 license.

A.4 FMRIPrep details for HCP working memory task

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.0.5 (Esteban et al. (2019); Esteban et al. (2018); RRID:SCR_016216), which is based on *Nipype* 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym], *ICBM 152 Nonlinear Asymmetrical template version 2009c* [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

Functional data preprocessing For each of the 14 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using `flirt` (FSL 5.0.9, Jenkinson and Smith, 2001) with the boundary-based registration (Greve and Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin6Asym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014).

The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

Copyright Waiver The above boilerplate text was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts *unchanged*. It is released under the CC0 license

B Supplement to results

B.1 Do basic statistical differences between HCP and Multi-task data affect transfer performance?

To better understand whether any basic differences in statistical properties, noise, or preprocessing between the HCP and Multi-task datasets affected the transfer performance of the pre-trained 3D-DeepLight variant, we performed a sequence of additional analyses.

We can immediately rule out basic differences in the temporal distribution of the voxel signals as we detrended and standardized the time series signal of each voxel within each fMRI run (to have a mean of 0 and unit variance; see section 4.1 of the main text). DeepLight further does not know about the temporal distribution of brain activity as it solely acts on the level of individual fMRI volumes. We therefore next probed the mean and standard deviation of voxel activities within each fMRI volume. We did not find any meaningful differences in the distribution of the volume means and standard deviations between the HCP and Multi-task datasets (see Appendix Fig. B.2).

We also tested whether other generic differences in noise between the HCP and Multi-task datasets affected transfer performance, by performing a confound correction of the Multi-task fMRI data, in which we regressed out variance related to the six motion correction parameters and three temporal and anatomical noise components resulting from fMRIPrep's CompCor method (for an overview, see Appendix Fig. B.4). Yet, the pre-trained model did not perform better when fine-tuned on the confound-corrected fMRI data than when fine-tuned on the fMRI data that was not confound corrected (for an overview of the training methods, see section 4.3 of the main text). 3D-DeepLight's final decoding accuracy on the confound-corrected data was 43.27%, thereby -2.5% worse than when applied to the uncorrected fMRI data ($t(5) = -4.65, P = 0.0056$; Appendix Fig. B.4).

Lastly, we also tested whether the transfer of the pre-trained model to the Multi-task data was affected by the different preprocessing that we applied to both datasets (we preprocessed the Multi-task dataset with fMRIPrep (Esteban et al., 2019), whereas the HCP uses an internal preprocessing pipeline; see section 4.1 of the main text). To this end, we downloaded the raw fMRI data of another 50 subjects in the HCP working memory task and also preprocessed these with fMRIPrep (for an overview of the preprocessing steps, see Appendix A.4). Interestingly, the pre-trained 3D-DeepLight variant again exhibited the advantages of transfer learning in this newly preprocessed fMRI dataset, by learning faster and achieving higher decoding accuracies than a model variant that was not pre-trained (see Appendix Fig. B.3). After training on the fMRI data of 20 subjects from this newly preprocessed dataset, the pre-trained model achieved a final decoding accuracy of 72.95% in the fMRI data of the remaining 30 subjects, while the model variant that was not pre-trained achieved a final decoding accuracy of 64.46% (i.e., -8.49% worse than the pre-trained model, $t(29) = -13.28, P < 0.0001$; see Appendix Fig. B.3; for an overview of the training methods, see section 4.3 of the main text).

Overall, we can therefore rule out that the transfer of the pre-trained model to the Multi-task dataset was affected by basic differences in the statistical properties, noise or preprocessing between the HCP and Multi-task datasets.

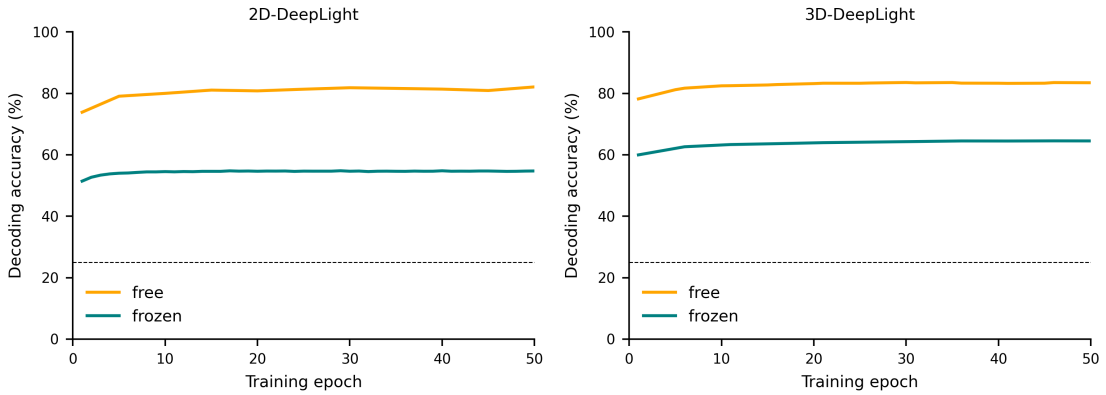

Figure B.1: Comparing two different fine-tuning approaches on the validation data of the HCP working memory task (see section 4.1.1 of the main text). We initialized the weights of two variants of each DeepLight architecture (left: 2D-DeepLight, right: 3D-DeepLight) to the weights of the pre-trained models (all except for the output layer, which now included four instead of 16 neurons; see section 4.2 of the main text for an overview of the architectures and Fig. 3 of the main text for an overview of the pre-trained model performance). We froze the pre-trained weights of one variant of each architecture during fine-tuning (depicted in green), while the other model variant was allowed to train all of its weights during fine-tuning (depicted in yellow) (see section 4.3 of the main text for an overview of the training procedures). Lines indicate decoding accuracy in the validation data as a function of the training epochs. Chance accuracy is indicated by the dashed horizontal line.

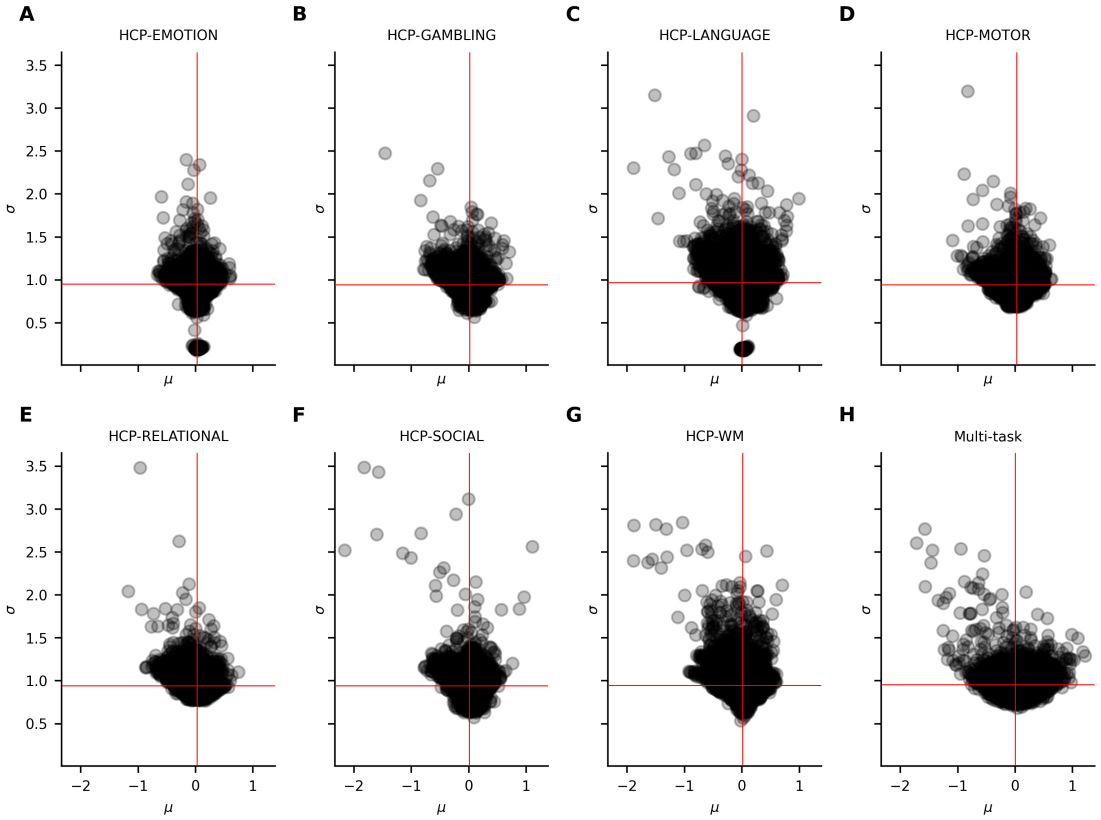

Figure B.2: Mean and standard deviation of voxel activities within each preprocessed fMRI volume in the validation datasets of the HCP experimental tasks (A-G) and Multi-task data (H) (for an overview of the datasets, see section 4.1 of the main text). Scatter points indicate individual fMRI volumes. Red lines indicate the mean over volumes.

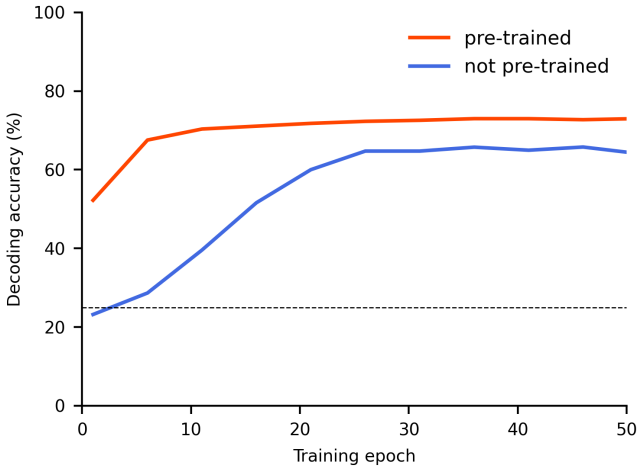

Figure B.3: Training decoding accuracy for a pre-trained (red) and not pre-trained (blue) 3D-DeepLight variant in the validation data of the HCP working memory task that was preprocessed with fMRIPrep (see Appendix B.1; see section 4.3 of the main text for an overview of the training procedures). An epoch was defined as an entire iteration over the training dataset. Lines indicate decoding accuracy. Chance accuracy is indicated by the dashed horizontal line.

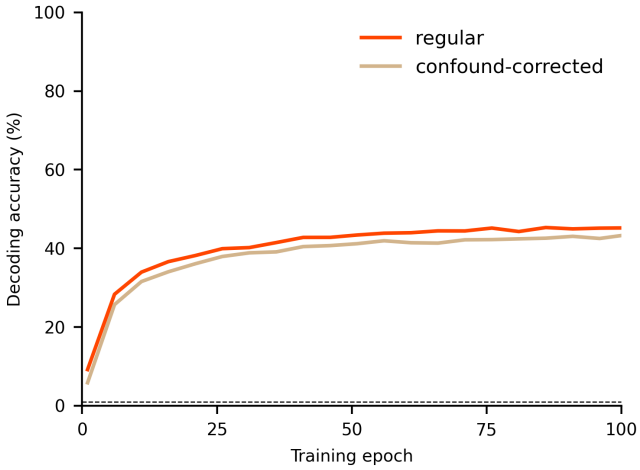

Figure B.4: Training decoding accuracy for the pre-trained 3D-DeepLight variant in two conditions: when it is fine-tuned on the regular fMRI data of the Multi-task dataset (red) or on a version that is corrected for basic noise confounds (tan). Specifically, we corrected the Mutli-task data for any variance resulting from the six parameters of basic motion correction, as well as the three temporal and anatomical noise components with the largest singular values resulting from fMRIPrep's CompCor method (for details on this method, see Behzadi et al. (2007)), by regressing their variance out of the time-series signal of each voxel (as implemented in Nilearns "signal.clean" function; Abraham et al., 2014). See section 4.3 of the main text for an overview of the training procedures. An epoch was defined as an entire iteration over the training dataset. Lines indicate decoding accuracy. Chance accuracy is indicated by the dashed horizontal line.

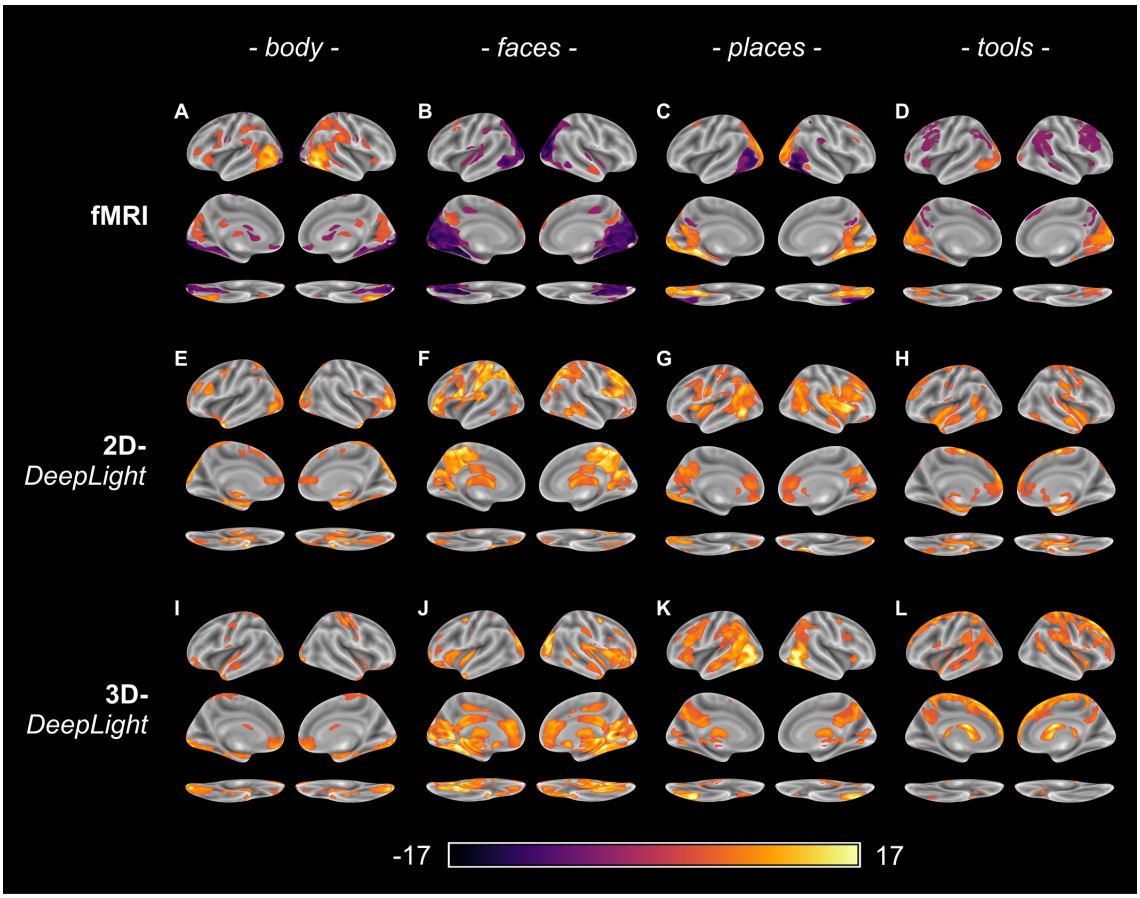

Figure B.5: Learned mappings between brain activity and cognitive states of the pre-trained DeepLight variants that were fine-tuned on the full training dataset of the HCP-WM experimental task (see section 2.3 of the main text). A-D: We first computed a standard two-stage GLM analysis (Holmes and Friston, 1998) of the fMRI data of the 50 subjects in the validation dataset of this task. E-L: We then also interpreted the decoding decisions of the 2D- (E-H) and 3D-DeepLight (I-L) variants for the same data. To identify the brain regions that each DeepLight variant associates most strongly with a cognitive state, we computed a similar two-stage GLM analysis of the resulting relevance data (restricting the resulting z-scores to only positive values). All GLM analyses were performed on parcellated brain data by the use of the dictionaries for functional modes (DiFuMo) atlas with 256 brain networks (Dadi et al., 2020) and computed separately for each experimental task by contrasting each cognitive state of the task against all other states of that task (for details on the GLM analysis, see Appendix A.2). All brain maps are thresholded at a false-discovery rate of 0.001 and projected onto the inflated cortical surface of the FsAverage template (Fischl, 2012). Brighter yellow values indicate larger z-scores.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: