CORTEX 168 (2023) 203–225

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Registered Report

Modeling face recognition in the predictive coding framework: A combined computational modeling and functional imaging study

Nestor Zaragoza-Jimenez [a,b,1], Hauke Niehaus [c,*,1], Ina Thome [a,b], Christoph Vogelbacher [a,b], Gabriele Ende [d], Inge Kamp-Becker [b,e], Dominik Endres [b,c,2] and Andreas Jansen [a,b,f,**,2]

[a] Laboratory for Multimodal Neuroimaging, Department of Psychiatry, University of Marburg, Germany
[b] Center for Mind, Brain, and Behavior (CMBB), University of Marburg and Justus Liebig University Giessen, Germany
[c] Theoretical Cognitive Science Lab, Department of Psychology, University of Marburg, Germany
[d] Department of Neuroimaging, Central Institute of Mental Health (CIMH), Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
[e] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University of Marburg, Germany
[f] Core-Facility Brainimaging, Faculty of Medicine, University of Marburg, Germany

ARTICLE INFO

Article history:
Protocol received 8 March 2021
Protocol approved 25 March 2022
Received 16 October 2022
Reviewed 5 December 2022
Revised 16 March 2023
Accepted 23 May 2023
Action editor Pia Rotshtein
Published online 26 July 2023

Keywords:
Face learning
Computational modeling
Predictive coding
Functional neuroimaging
Functional magnetic resonance

ABSTRACT

The learning of new facial identities and the recognition of familiar faces are crucial processes for social interactions. Recently, a combined computational modeling and functional magnetic resonance imaging (fMRI) study used predictive coding as a biologically plausible framework to model face identity learning and to relate specific model parameters with brain activity (Apps and Tsakiris, Nat Commun 4, 2698, 2013). On the one hand, it was shown that behavioral responses on a two-option face recognition task could be predicted by the level of contextual and facial familiarity in a computational model derived from predictive-coding principles. On the other hand, brain activity in specific brain regions was associated with these parameters. More specifically, brain activity in the superior temporal sulcus (STS) varied with contextual familiarity, whereas activity in the fusiform face area (FFA) covaried with the prediction error parameter that updated facial familiarity.

Literature combining fMRI assessments and computational modeling in humans still needs to be expanded. Furthermore, prior results are largely not replicated. The present study was, therefore, specifically set up to replicate these previous findings. Our results support the original findings in two critical aspects. First, on a group level, the behavioral

* *Corresponding author*. Department of Psychiatry, University of Marburg Rudolf-Bultmann-Straße 8, 35039 Marburg, Germany
** *Corresponding author*. Laboratory for Multimodal Neuroimaging, Department of Psychiatry, University of Marburg, Germany
E-mail addresses: niehaush@staff.uni-marburg.de (H. Niehaus), andreas.jansen@staff.uni-marburg.de (A. Jansen).
[1] Contributed equally.
[2] Shared senior authorship.
https://doi.org/10.1016/j.cortex.2023.05.021

imaging
fMRI
Core system
FFA
pSTS

responses were modeled best by the same computational model reported by the original authors. Second, we showed that estimates of these model parameters covary with brain activity in specific, face-sensitive brain regions. Our results thus provide further evidence that the functional properties of the face perception network conform to central principles of predictive coding. However, our study yielded diverging findings on specific computational model parameters reflected in brain activity. On the one hand, we did not find any evidence of a computational involvement of the STS. On the other hand, our results showed that activity in the right FFA was associated with multiple computational model parameters. Our data do not provide evidence for functional segregation between particular face-sensitive brain regions, as previously proposed.

1. Introduction

The recognition of familiar faces is crucial for appropriate social interactions. Recognizing a familiar face comprises the recognition of someone's visual appearance and entails the spontaneous retrieval of autobiographical information and the triggering of appropriate emotional responses. At the neural level in humans, the recognition of familiar faces is associated with a distributed hierarchical cortical brain network. This network is often divided into a *core system* and an *extended system* (Duchaine & Yovel, 2015; Haxby, Hoffman, & Gobbini, 2000; Said, Moore, Engell, Todorov, & Haxby, 2010). The *core system* encodes the visual appearance of a face, and it consists of at least three bilateral, typically right-lateralized regions: the occipital face area (OFA) in the inferior occipital gyrus, the fusiform face area (FFA) in the lateral fusiform gyrus, and the posterior superior temporal sulcus (pSTS). The OFA, as the hierarchically first stage of the face-processing network, is typically associated with the analysis of invariant facial features like eyes or mouth and the decision of whether an object is a face or not (Pitcher, Walsh, & Duchaine, 2011). The FFA mainly processes abstract high-level features of faces and therefore plays a key role in recognition of familiar identities (Kanwisher, McDermott, & Chun, 1997; Kanwisher & Yovel, 2006), although specificity and invariance of FFA activation/representation are still debated (see, e.g., Tsantani et al., 2021 or Burns, Arnold, & Bukach, 2019). The pSTS is involved in perceiving changeable features of faces, such as expression and eye gaze. It has also been associated with recognizing familiar visual appearance linked to the perception of dynamic components characteristic of an individual (O'Toole, Roark, & Abdi, 2002). The *extended system* is involved in extracting further information that a face can convey (Haxby et al., 2000). Diverse cognitive processes related to our ability to recognize familiar faces have been linked to multiple brain regions. Anterior temporal areas are associated with the representation of semantic and biographical information, while the precuneus is related to the retrieval of episodic memories. The amygdala and the insula, areas involved in the representation of emotions, have been related to the processing of emotional responses associated with the recognition of specific faces. Familiar faces also recruit "Theory of Mind" areas, such as the temporoparietal junction (TPJ), to a more significant extent than unfamiliar faces because they are more strongly associated with personal knowledge (Gobbini et al., 2007). The computational processes that underlie the learning of facial identities, i.e., the transition of faces from unfamiliar to familiar, are still poorly understood in humans.

In theory, these processes could be modeled within the meta-framework of the *Hierarchical Mechanistic Mind* (HMM, Badcock, Friston, & Ramstead, 2019; Badcock, Friston, Ramstead, Ploeger, & Hohwy, 2019), subsuming the recent neuro-theoretic frameworks of the Bayesian brain, predictive coding, and active inference under the free-energy principle. Following the HMM meta-framework, predictive coding (PC) is a neurobiologically plausible algorithmic and implementational scheme describing how inference over the latent causes of sensory input might be implemented in humans and other mammals (Clark, 2013; Hohwy, 2020; Keller & Mrsic-Flogel, 2018). PC rests on the proposition that the brain implements a hierarchical probabilistic generative model of the world, which constantly maintains a set of hypotheses about the expected sensory input, thereby coding the generative mechanisms causing the sensory input under consideration of context and previous experience. This generative model uses these hypotheses to predict each level of the computational hierarchy regarding how likely a sensory input is to be expected under a given generative hypothesis. The predictions derived from each of the hypotheses maintained by the internal model and sensory data are compared to calculate a *prediction error*. The brain then attempts to infer the causes of the sensory input by iteratively reducing this prediction error either by an update of the internal generative model (i.e., "explaining away" the prediction error/updating hypotheses and predictions) and/or by choosing appropriate actions (e.g., deciding to acquire more sensory data or to modify the environment to conform to the predictions). During the last decade, PC has been examined in various studies to describe the functional brain organization, and canonical cortical circuits have been proposed (Keller & Mrsic-Flogel, 2018). The PC framework has now been used in a wide range of neuroimaging studies, from visual perception to higher cognitive processes, e.g., language (Brodski-Guerniero et al., 2017; Lewis & Bastiaansen, 2015; Rauss, Schwartz, & Pourtois, 2011; Shipp, Adams, & Friston, 2013) or memory (Barron, Auksztulewicz, & Friston, 2020; Frank & Kafkas, 2021). In summary, it can be

concluded that PC-consistent approaches for studying multiple neuroscientific phenomena have emerged as plausible and useful process models (Teufel & Fletcher, 2020).

Consequently, it has also been proposed that examining the hierarchical nature of facial processing under a PC perspective might help disentangle current issues with apparently contrary results regarding the specificity and hierarchy within the core and extended face networks. Indeed, in the macaque face-processing system, preliminary evidence has been reported for neuronal computations consistent with PC, which helped advance the field (Freiwald, 2020; Issa, Cadieu, & DiCarlo, 2018; Meyer & Olson, 2011; Schwiedrzik & Freiwald, 2017). However, some of the prior literature yielded negative results regarding PC consistency in macaque facial processing studies (e.g., Kaliukhovich & Vogels, 2011; Vinken, Beeck, & Vogels, 2018; Vogels, 2016).

In humans, using task-based functional magnetic resonance imaging (fMRI), a large body of literature also indicates the compatibility of activity in relevant cortical face-processing areas with predictive coding accounts (Freiwald, 2020; Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008; Summerfield & Koechlin, 2008; Summerfield & Lange, 2014; Trapp, Schweinberger, Hayward, & Kovács, 2018). Notably, a previous study modeled face-stimulus-induced FFA activation in humans via PC-consistent computational mechanisms (Egner, Monti, & Summerfield, 2010). Their study demonstrated that FFA activity in a cued object/face discrimination task was best modeled by a group-level (fixed-effects) regression model using expectation and error-related responses as regressors.

Following the previously mentioned results, in this work, we were focused specifically on fMRI and computational modeling results for a PC-consistent neural hierarchy in the human face processing system. Importantly, human results are available, including computational modeling-informed fMRI approaches. Here, we decided to replicate a multimodal study by Apps and Tsakiris (2013) that successfully employed a PC-consistent behavioral auto-regressive (time-series) computational model in combination with task-based fMRI. To our knowledge, this was the first successful model of this kind. The original authors modeled informed fMRI analysis with a facial identity learning paradigm to examine the in-principle compatibility of neural activity in the face-processing framework with PC-derived computational hypotheses. Apps and Tsakiris (2013) expanded upon the prior literature with new computational models formalizing multiple competing hypotheses regarding face identity learning. Firstly, their computational modeling approach improved upon the model by Egner et al. (2010) in two important aspects: (a) it allows a subject-level model-fitting procedure, thereby enabling random-effects analyses of estimated individual model parameters (e.g., learning rates) and (b) estimation of trial-by-trial values of dynamic latent variables (e.g., prediction errors) thought to be represented by the subject during task performance.

In this work, we performed a replication of the effects of Apps and Tsakiris (2013) that is akin to a direct replication. We followed in our approach the definition of a direct replication as *"a study that attempts to recreate the critical elements (e.g., samples, procedures, and measures) of an original study"* but *"[…] does not have to duplicate all aspects of an original study. Rather it must only duplicate those elements that are believed necessary for producing the original effect."* (Zwaan, Etz, Lucas, & Donnellan, 2017, p. 3, p. 3). To identify replication candidates, we performed a literature overview, and the study by Apps and Tsakiris (2013) emerged as the replication candidate with the highest expected utility upon replication (Isager, et al., 2021a, b) in its field using this specific methodology and theoretical background. The literature overview was performed over the years 2000–2021 with the key topics 'face processing,' 'computational modeling,' 'predictive processing,' and 'fMRI,' including their synonyms (see Supplementary material/Appendix A and B for exact keywords, parameters, and full results; see Table 1 for study details of the two relevant replication candidates after screening). Furthermore, we estimated the expected utility gain of replication specific for each of the replication candidates via the replication value (RV_{Cn}; Isager, et al., 2021a, b).

The results of our literature overview were further corroborated by Trapp et al. (2018), where the authors highlight abundant attempts in the literature to leverage the well-researched neural facial processing system to empirically test PC hypotheses without reference to existing cognitive models. The authors contrasted this with the relative scarcity of literature that attempts to test the general compatibility of cognitive face processing models with PC principles or to reformulate current cognitive models in a predictive framework. Contrary to this latter trend, Trapp et al. (2018) singularly identified the study by Apps and Tsakiris (2013) as the first successful attempt to connect high-level face recognition processes with PC principles in humans.

Within the PC framework, Apps and Tsakiris (2013) modeled face identification as a function of *contextual information* and *stimulus familiarity*. In PC, contextual information is used to shape the prediction of, e.g., how familiar an upcoming stimulus will likely be, thereby modulating the processing of that stimulus. Context can be operationalized, for instance, by the familiarity with objects that are presented before the presentation of a target stimulus. Stimulus familiarity naturally also influences the recognition process. To operationalize both influences on the face recognition process, Apps and Tsakiris (2013) constructed a cognitive task where subjects were repeatedly shown computer-generated faces presented from different viewpoints (front on or at a 30° angle from the right or left). *Stimulus familiarity* was operationalized by the number of times a specific facial identity was presented before a given trial overall and *contextual familiarity* was operationalized by the general level of *stimulus familiarity* in the immediate history of a given trial, i.e., the number of 'familiar' faces previously presented. All face stimuli were unknown at the beginning of the experiment to the subjects and some facial identities were pseudo-randomly shown more often than others. As a result, familiarity and contextual information were varied systematically over the stimulus presentation. The participants were required to indicate in a two-option forced-choice task whether they had seen that person before or not, regardless of the specific viewpoint. Importantly, Apps and Tsakiris (2013) developed a set of computational models coding multiple computational hypotheses consistent and inconsistent with the PC

Table 1 – Two remaining relevant replication candidates after screening with relevant study details. A higher replication value indicates a higher expected utility upon replication. See Supplementary material/Appendix A for exact formula of replication value.

	Year of publication	Number of citation	Age in years	Yearly citation rate (2021)	Effective sample size	Replication value (RV_{Cn})
Stefanics et al.	2019	6	2	3.0	35	.58
Apps and Tsakiris	2013	20	8	2.5	16	1.66

framework. Their models allowed the computation of unique/idiosyncratic parameters like learning rates and trial-by-trial estimates of hidden quantities like prediction errors. Their winning PC-consistent model was able to make predictions about how each face presentation increased the probability that the same face will be recognized and how such effects might be modulated by contextual information.

On a behavioral level, Apps and Tsakiris (2013) demonstrated that subjects showed a clear learning effect, categorizing the faces as familiar the more times a facial identity had been presented. In their behavioral task, they showed that responses on the face recognition task could be predicted by the level of both contextual familiarity and stimulus familiarity irrespective of viewed perspective. Using fMRI, they further showed that activity in the FFA varied with the trial-by-trial prediction error parameter that updated stimulus familiarity. Activity changes in the pSTS were associated with contextual familiarity. Taken together, their results characterized, both at the behavioral and at the brain level, key computational mechanisms underlying the perceptual learning of faces. The examination of facial processing with combined computational modeling and functional neuroimaging is fruitful from a basic neuroscientific perspective but is unfortunately rare and not replicated in the current literature.

In summary, the present study aimed at replicating the central findings and methodology of Klicken oder tippen Sie hier, um Text einzugeben (2013). Our replication goals were focused on a direct replication of the original findings. However, as some methodological parameters were changed in our design (compared to the design of the original authors), this replication attempt will only approximate an idealized direct replication. The central findings we attempted to replicate can be summarized as follows.

- Hypothesis 1: Behavioral responses on a two-option face recognition task can be modeled best (highest approximate Bayes factor, see section 2.4.8) by a computational model consistent predictive-coding principles that assume combined effects of contextual familiarity and facial familiarity irrespective of viewing angle ('view-independence and context'-model). A detailed mathematical specification is given in the Methods section 2.4.8.
- Hypothesis 2: Activity in the right pSTS is correlated with the trial-by-trial contextual familiarity estimated by the computational model described in Hypothesis 1.
- Hypothesis 3: Activity in both the right and the left FFA is correlated with the trial-by-trial prediction error estimate that updates facial familiarity in the winning model described in Hypothesis 1.

The methods section will describe if and why changes were introduced to the study design in comparison to the original study by Apps and Tsakiris (2013). Stage 1 preregistration is available under the link https://doi.org/10.17605/OSF.IO/A8VU7.

2. Methods

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

2.1. *Subjects*

Thirty-four subjects (13 male, 21 female; mean age 23.8 years ± 2.6 years) were included in the study. Inclusion criteria were age between 18 and 30 years, right-handedness according to the Edinburgh handedness inventory (Oldfield, 1971) with a cut-off at +30, German as a native language, and normal or corrected-to-normal vision. Exclusion criteria were COVID-19-like symptoms, alcohol or drug abuse, past or present psychiatric or neurological disorders according to ICD-10, reading or spelling disorders, and disturbances of color vision. During pre-registration, we established several criteria according to which subjects had to be excluded. Two participants were excluded because of bad MRI data quality. One subject was excluded because it did not reach a level of correct face recognition of at least 60% averaged over all stimuli at the end of the experiment (see 2.3.1, behavioral data quality, for details). Thirty-one subjects (11 male, 20 female; mean age 23.7 years ± 2.6 years) were therefore included in the final data analysis. Before the experiment, written informed consent was obtained following the Declaration of Helsinki. 10 € compensation was offered for participation. The study design was approved by the local ethics committee of the Medical Faculty of the University of Marburg.

2.2. *Power analyses*

The required number of subjects was calculated by a power analysis. We performed separate power analyses for all three central aspects of Apps and Tsakiris (2013): the behavioral results, the fMRI results, and the computational model selection. A minimum required sample size of 26 was determined by the fMRI analyses; the behavioral and computational model selection results yielded a minimum required sample size of three ($\alpha = 0.02$) and ten, respectively, for the power of > 90%.

2.2.1. Behavioral power analysis

The power analysis for the behavioral learning effect was performed using the data extracted from the original study (see Table 2). The original study reports a separate analysis to estimate a learning effect via one-sample t-tests. The learning effect was defined as an increased probability of answering *yes* (i.e., declaring to recognize a face) with an increasing number of prior presentations, irrespective of whether the answer was correct or not. We interpreted this analysis on a group level by comparing the overall mean percentage of stimuli that yielded *yes* responses at their first three and last three presentations. Here we utilized the observed t-value and the sample size described by Apps and Tsakiris (2013) to estimate an effect size and compute a power analysis using the function 'pingouin.power_ttest' of the python package Pingouin (Vallat, 2018).

In the first step, we converted the t-value obtained by the t-test of the original study to an effect size (i.e., Cohen's d) using the reported sample size of $N = 16$. Next, to estimate the sample size needed for replication, we specified a power analysis with the converted effect size, a *power* of .9, and an *alpha level* of .02.

2.2.2. fMRI power analysis

Apps and Tsakiris (2013) showed that activity in the right pSTS varied with contextual familiarity C_t, whereas activity in both the left and the right FFA co-varied with the prediction error parameter ε that updated face familiarity. Our power analysis specifically aimed to determine the number of subjects needed to replicate these findings.

The required number of subjects for the fMRI analysis was calculated by a power analysis using the software G*Power (version 3.1, Faul, Erdfelder, Buchner, & Lang, 2009). The statistical power was set to .90, according to the *Registered Report Guidelines*. The alpha error probability was set to $\alpha = .02$, again following the *Registered Report Guidelines*. Since we had to correct the number of main analyses, we divided this number by 3, resulting in a final threshold of $\alpha = .0067$. The expected effect size (i.e., Cohen's d) was estimated from results presented in the original study (i.e., from the averaged parameter estimates and the respective standard errors, see Fig. 3, p. 6 in Apps and Tsakiris (2013)). This yielded, averaged across all face regions, an effect size estimate of $d = .788$. Using these numbers, the power analysis (based on a one-sample t-test) yielded a required sample size of $n = 26$ subjects (Please note that we erroneously wrote in the original version of the manuscript (i.e., at stage 1 of the registered report) that we had to include $n = 27$ subjects.).

2.2.3. Computational modeling power analysis

To estimate the minimum sample size required for robust results of the computational model selection, we employed a generative model-based power analysis based on our pilot sample of 3 subjects. Following Gluth and Jarecki (2019), we extended the statistical concept of power as the probability of our model selection approach to correctly identify the data-generating mechanism producing the observed behavior/data with a probability of $\geq 90\%$ on a group level. More specifically, this translates to a relative posterior model probability of our winning model of ≥ 0.9. The posterior model probability of the proposed winning model was calculated from the group-level Bayesian information criterion (BIC) sums of the models via formula (14) of the main text. We utilized the individually fitted *view-independent model with contextual influence* and the *view-dependent model without contextual influence* from our pilot data to generate new synthetic samples of arbitrary plausible sample size n^*. To estimate the required sample size, we simulated two opposite data-generating scenarios, as summarized in Table 3.

First, we created a *best-case scenario* where the subject-data is produced by a generative mechanism involving view-independent representation of face stimuli with contextual influence. Note that this model was also proposed to be the winning model in both scenarios, as it was reported as the winning model by Apps and Tsakiris (2013). Here, we estimated a required sample size that ensured a $\geq$90% posterior model probability of this model.

Secondly, we simulated a *worst-case scenario* where subject-data is generated by a generative mechanism only involving view-dependent representation of face stimuli without any contextual influence. Further, we estimated a required sample size that ensured a posterior model probability of $\leq$10% (1-best-case scenario) for the view-independent proposed winning model with contextual influence. Since the synthetic data were generated by a different generative mechanism as in the proposed winning model, and the context-dependent view-independent model should therefore have a low ($\leq$10%) posterior model probability.

Points 1, 2, and 3 of the following procedure were done in each scenario described in Table 3. The entire procedure, i.e., simulating each scenario, was repeated for each time point.

First, we generated a large set N^* of artificial subject data from each generative mechanism/model with a respective total size of $|N^*| = 500$. This data set was generated based on the fitted models from our pilot data.

Second, for each relevant sample-size n^* and for each generative mechanism (cf. Table 3), we drew 50 random sub-samples with replacement of size n^* from N^* and fitted the proposed winning model to each sub-sample, and calculated the BIC values per model. However,we assumed the view-independent and context components as the data generating process. In other words, the data were generated by a 'best-case scenario' and a 'worst-case scenario' with respect to the

Table 2 – Power and results reported by Apps and Tsakiris (2013) regarding the learning effect, parameters, and results of our power analysis based on the probability of 'yes' answers and observed power in our pilot sample.

		t-value	Cohen's d	N	Power	Alpha/p-value
Apps and Tsakiris (2013)		14.661	n.a.	16 (reported)	>.999	$p < .0001$
Our analysis	Power analysis	n.a.	7.33	~3 (required for replication)	.9	$\alpha = .020$
	Our pilot data	20.23	5.14	3 (reported)	.98	$P = .002$

Table 3 – Summary of the two generative mechanisms, which model is assumed to be the winning model and power criteria.

Simulated scenario	Generative-mechanism of the data	Proposed winning model	Power criterion (posterior model probability)
Best-case	*view-independent representation, contextual influence*	*view-independent model with contextual influence*	≥ 90%
Worst-case	*view-dependent representation, no contextual influence*	*view-independent model with contextual influence*	≤10%

underlying generative mechanism. The corresponding proposed winning model should have the highest posterior model probability (Table 3) in the 'best-case scenario' and the lowest in 'worst-case scenario' consistently over subsamples.

Third, we calculated the group-specific posterior model probability of the proposed winning model (view-independent and context) based on the per-model sum of the BIC over all n^* artificial subjects in each sub-sample.

Fourth, the sample size n^* was increased until (1) the mean posterior probability of the proposed winning model (*view-independent model with contextual influence*) exceeded 90% while the data was generated assuming this generative mechanism and until (2) the mean posterior probability of the proposed winning model did not exceed 10% while the data was generated by the view-dependent model without assuming contextual influence.

Finally, we decided upon the minimum sample size of $N = 10$ as it consistently yielded a posterior model probability within the specified parameters with a negligible variation (see Fig. 1) and a 95% CI lower-bound of >.999.

The detailed results of the power analysis can be found in Table 4.

2.3. *Experimental paradigms*

Subjects performed two tasks. The first task was a face identity learning task. In this task, subjects were repeatedly shown faces whose identity was initially unknown. They had to judge

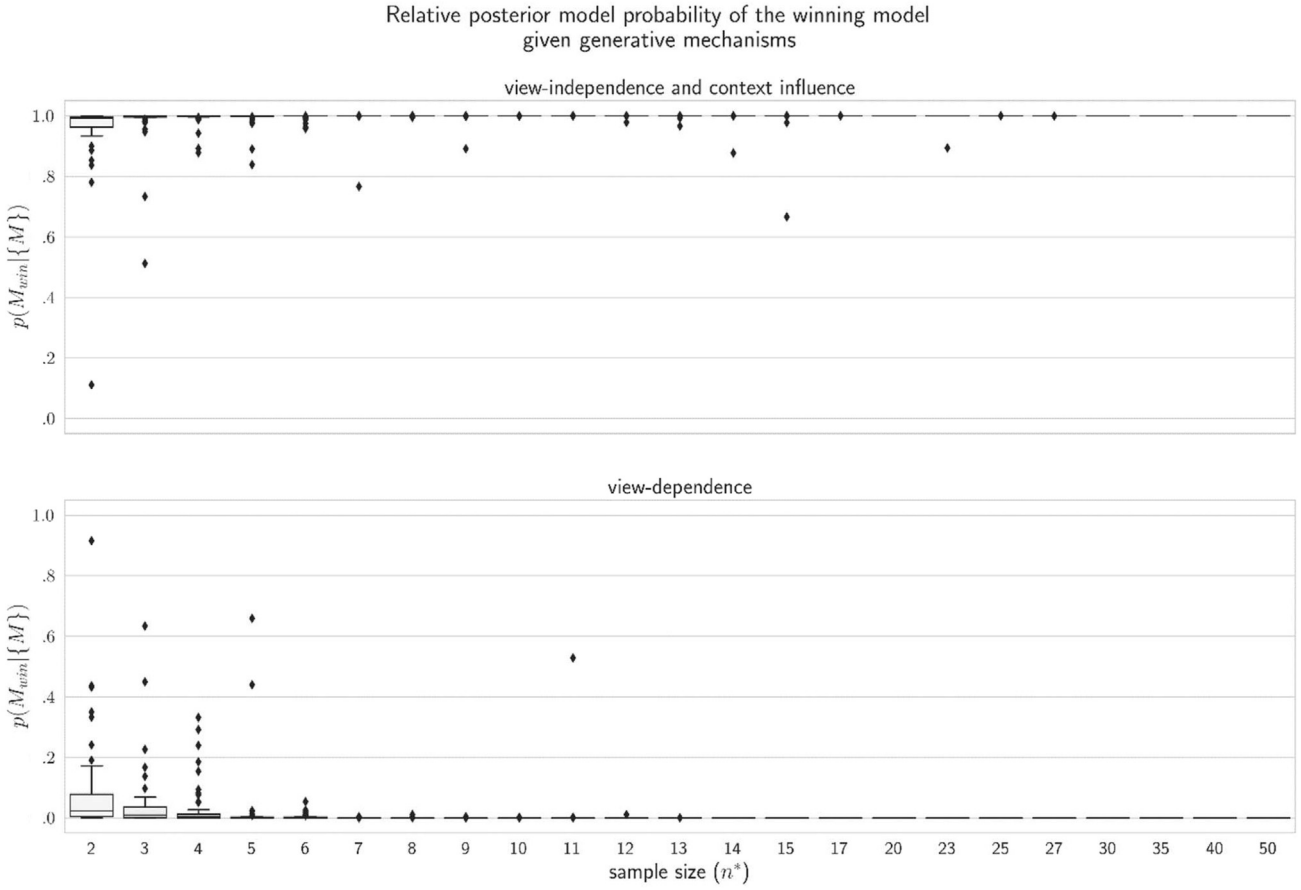

Fig. 1 – Boxplots of the power analysis for computational modeling approach. Datapoints within each sample size describe a result for a subsample. The Y-axis of the subplots shows the posterior model probability of the proposed winning model 'view-independent model with contextual influence.' The X-axis of the subplots shows the sample sizes n^* used for power estimation.

Table 4 – Detailed results from the power analysis for a sample size of $n^* = 10$. The proposed winning model was, in both scenarios, the 'view-independent model with contextual influence.

Simulated scenario	Selected n^*	Posterior model probability (proposed winning model)	95% CI for selected n^* [lower bound, upper bound]	Criterion
Best-case	10	>.999	[.999.1]	≥ .9 (90%)
Worst-case	10	<.001	[<.001, <.001]	≤ .1 (10%)

whether a face had been presented before or not. The ground truth of the task used to generate the stimulus order over time consisted of the *stimulus familiarity* (i.e., the number of times a specific face stimulus has been presented before) and the *contextual familiarity* (i.e., the average familiarity of the previously presented faces). Both parameters of the task were systematically varied over the course of the trials. The second task was a face localizer task in which blocks of faces and non-face stimuli, respectively, were presented. This task allowed us to visualize individual face-sensitive brain areas (OFA, FFA, pSTS), making it possible to assess brain activity changes in the first task in independently selected regions of interest (ROIs). In the following, we will describe the face identity learning task in detail. The description of the face localizer task can be found in the supplementary material/ Appendix J.

2.3.1. *Face identity learning task*

2.3.1.1. Stimulus material. We created face stimuli as close as possible to the study of Apps and Tsakiris (2013). The stimuli were taken from the database of the Social Perception Lab, Princeton (http://tlab.princeton.edu/databases/secretdatabaseportal/). More specifically, we used the "300 random faces" data set (Todorov, Said, Engell, & Oosterhof, 2008), consisting of 300 faces that were randomly generated using the FaceGen software package (version 3.1, https://facegen.com/modeller.htm). These faces were rated on nine trait dimensions: attractiveness ($n = 35$), likeability ($n = 32$), trustworthiness ($n = 29$), competence ($n = 44$), extroversion ($n = 33$), dominance ($n = 23$), meanness ($n = 27$), frightening ($n = 28$), and threatening ($n = 27$). We selected 24 faces within one standard deviation of the mean on each of the mentioned trait dimensions, i.e., the faces were close to the mean on all traits (see Fig. 2 for an example).

Face stimuli were presented centrally on a uniform gray background, with 15° height and 12° widths of the visual angle. As face identity recognition is typically guided by visual features beyond facial structure, such as skin color, we used color face stimuli. All faces from the database were shown from the front. To investigate face identity recognition independent of the point of view, we generated two face stimuli for each individual using the FaceGen software. The face was turned 30° to the left and right in these stimuli, respectively. Participants were required to recognize faces from novel viewpoints, even if they had not seen that face from this specific viewpoint before.

2.3.1.2. Experimental design. Subjects were repeatedly presented with computer-generated faces ("trials"). These faces were shown from different viewpoints. After each presentation, the subjects were required to indicate whether they had

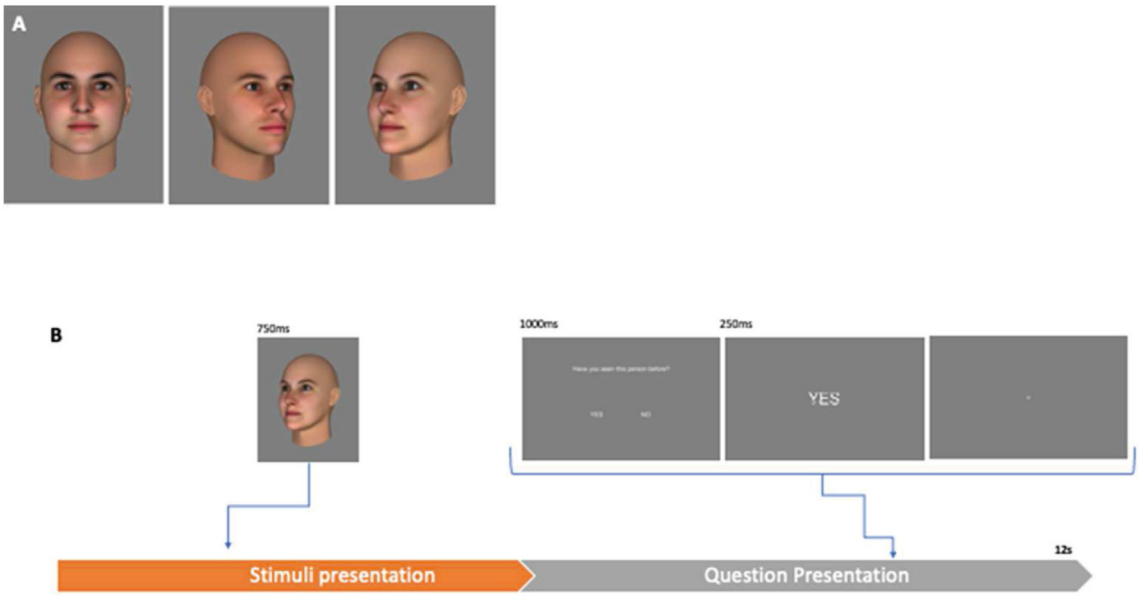

Fig. 2 – (A) Example of computed generated face identities. The images are displayed in the three different points of view we used during the experiment (B) Experimental design, trial example. A face identity was initially presented, followed by a cue question, short feedback, and a jitter randomly distributed in 12 sec intervals. The face stimuli were randomly presented from 0 to 5.25s. The question was also randomly presented within a 6–12s interval.

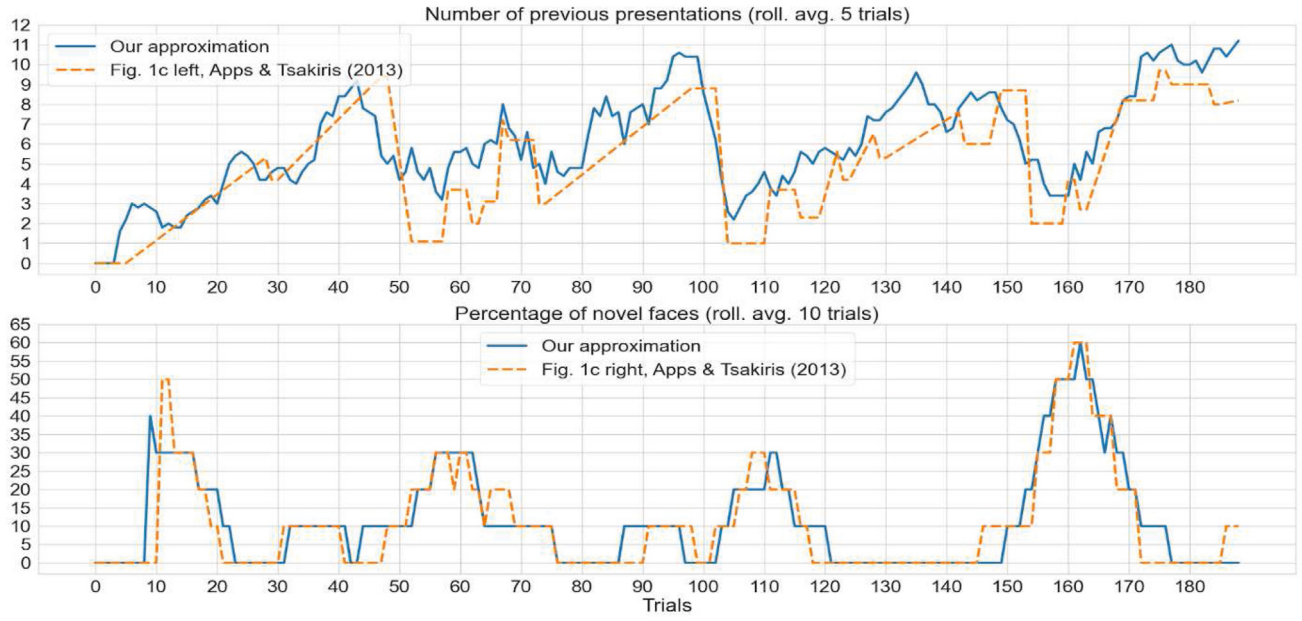

Fig. 3 – Ground truth data was used for the generation of our stimulus sequence estimated iteratively due to missing f information in the original study (missing Fig. 4 in suppl. Material of the original study). The six-block structure mentioned in the original study could not be reproduced. The upper figure shows the (graphically) extracted data from Fig. 1 c left from the original study versus our best approximation. Here, the rolling average of our trial sequence was calculated with respect to the rolling average number of previous stimulus presentations. The original authors report manipulating this value deliberately, as it is centrally important for contextual familiarity calculation. The lower figure shows the data from Fig. 1 c, right from the original study. Here, the rolling average number of new faces in the last ten trials is depicted.

seen that person before or not in a two-option forced-choice task, regardless of the specific viewpoint. In most aspects (e.g., regarding the timing of the onsets), our experimental design was as close as possible to the original study design of Apps and Tsakiris (2013). Of note, however, some experimental details (e.g., the operationalization of contextual familiarity) were not evident in the original paper. We numerically reconstructed the experimental design from the available graphical data in the article document, as detailed in Fig. 3.

The experiment consisted of 189 trials. Before the first and last trials, subjects were presented with a white screen for 7.250 sec. Each trial lasted 12 sec. First, a face was shown for 750 ms. It was shown in the first half of the trial, i.e., the onset of the presentation was between 0 sec and 5.250 sec. The onset was randomly varied across trials and subjects. Second, the question "Do you recognize the person?" was displayed in the center of the screen for 1000 ms. During this time, the participants were allowed to answer via a button press. Below the question, the words "yes" and "no" were presented to the right and left of the center of the screen, respectively. To prevent participants from preparing a motor response already at the time of the face presentation, the "yes" and "no" stimuli were pseudo-randomly positioned on the right and left of the screen. The response given by the participant was displayed on the screen for 250 ms as feedback directly after the presentation of the question. The feedback ensured the participants that their responses were recorded. If the participant did not respond to the trigger cue within a time window of 1000 ms, the word "missed" was displayed as feedback on the screen. Question and feedback were shown in the second half of the trial, i.e., the onset of the presentation of the question was between 6 sec and 10.750 sec. The onset was randomly varied across trials and subjects.

Twenty-four different face identities were shown in the experiment. All faces were unknown at the beginning of the experiment. Fifteen face identities were shown several times and were thus learned. These faces were presented from different viewpoints (front on or at a 30° angle from the right or left). Each of these faces was presented four times from each viewpoint, i.e., 12 times in total. Nine face identities were shown only once. Three of these faces were presented front on, three at a 30° angle from the right and three at a 30° angle from the left. In total, 189 face stimuli were shown.

Face stimuli were ordered so that the identities of the faces became gradually more familiar to the participants throughout the experiment. For this design, the order of the face stimuli was partly pseudorandomized and partly controlled. On the one hand, both the order of identities and the order with which each identity was presented from the different viewpoints were pseudorandomized. Subjects, therefore, were not able to make any predictions about the upcoming stimulus besides the prediction based on contextual familiarity. On the other hand, contextual familiarity during the presentation of a stimulus was systematically controlled. Contextual familiarity was operationalized by the number of previous presentations of a familiar face. By manipulating the average number of times a face identity was seen before, we were thus able to control the contextual

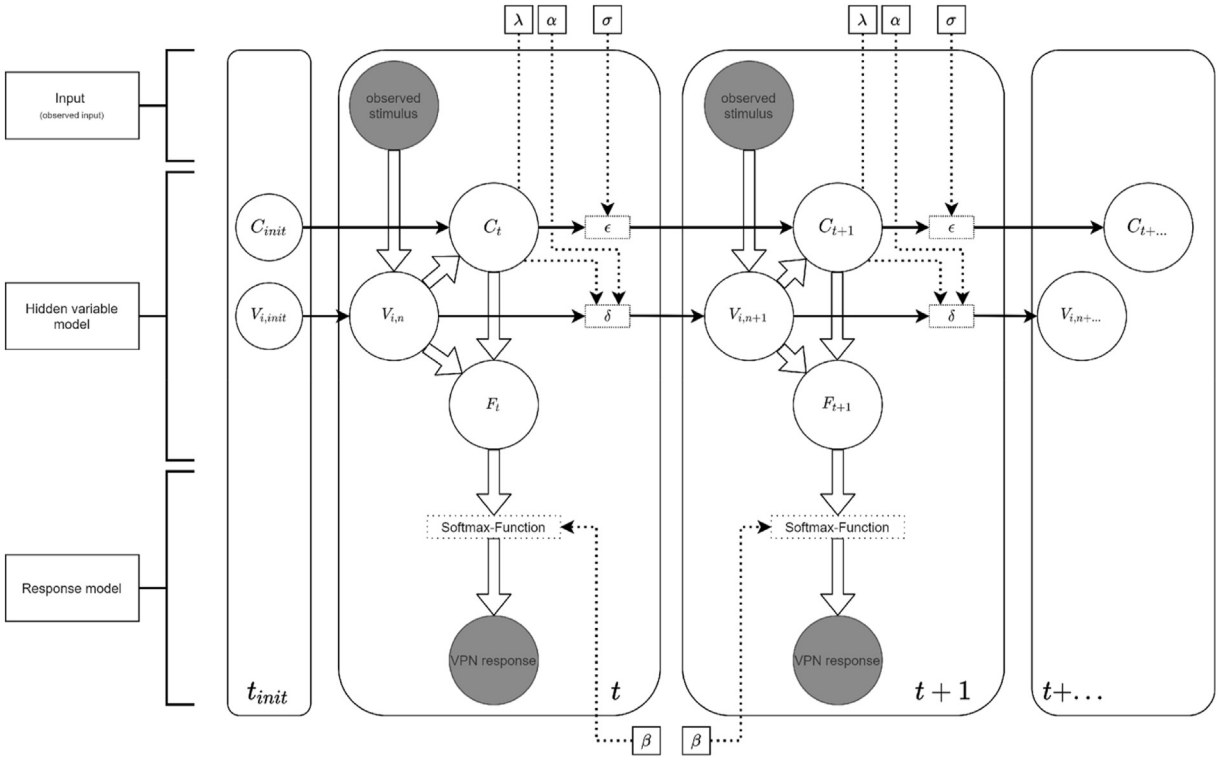

Fig. 4 – Graphical depiction of the 'view-independent and context-dependent model' (Model 1, proposed winning model). Squares indicated idiosyncratic parameters estimated per subject. See Table 6 for additional details. Dotted rectangles denote the trial-by-trial updates to hidden variables, including prediction errors. Large white circles denote the hidden variables of this model. Note that hidden variables are different for each model, see Table 7 for details. Filled circles denote the observed data by each subject and the corresponding action. Arrows denote general mathematical dependencies. Index *t* refers to trials, and index *init* refers to the initialization values.

familiarity without losing the randomization effect for viewing angle and identity. This is detailed in the upper part of Fig. 3. A further feature that was controlled was the presentation of the nine face stimuli that were presented only once. They were shown in particular in the later stages of the experiment so that subjects were not able to perform the task more accurately by increasing the number of "yes" responses. All blocks, therefore, contained face identities that had not been presented before or had been presented only a small number of times during the experiment.

Before the fMRI measurements, all participants took part in a training session. This session consisted of a short version of the task that the participants performed inside the MR scanner. The face identities in the training session differed from the face identities in the scanning session but were taken from the same database from which the original stimuli were created. Before the scanning session, participants were informed that they would be 'learning' a completely novel set of face identities.

2.3.1.3. Task performance. Apart from computational modeling and fMRI analyses, one of the main goals of this work was the replication of behavioral results observed in Apps and Tsakiris (2013). Therefore, the following task performance metrics were calculated per subject.

- Total errors made (false remembering & false not remembering)
- Percentage of correct/wrong answers from all trials
- Answers missed

Following the original study, we performed another series of *t*-tests to ensure learning of the face identities. We calculated the proportion of 'yes' answers for the first and last three presentations (i.e., 1st.2nd, 3rd vs. 10th,11th, 12th) of each stimulus averaging over all subjects. Assuming adequate learning of the face identities, the proportion of 'yes' answers was assumed to be significantly higher in the last three presentations than in the first three.

2.3.1.4. Behavioral data quality. During pre-registration, we established several criteria according to which subjects had to be excluded. The first criterion was based on the behavioral quality. To ensure behavioral data quality, subjects had to reach a level of correct face recognition of at least 60% averaged over all stimuli at the end of the experiment. More specifically, the percentage of correct responses was averaged for the last three presentations of each stimulus for stimuli that were presented more than once. This stimulus-specific average was then, in turn, averaged again over all relevant stimuli resulting in a single average percentage indicating the

overall highest mean recognition performance reached by a subject at the end of the experiment. One subject did not reach this threshold and was therefore not included in the final analysis and excluded from any further assessments.

2.4. Computational modeling

Apps and Tsakiris (2013) developed multiple computational models to test several hypotheses pertaining to computational processes underlying face identity learning. These models can be described as a non-hierarchical implementation of learning principles also found in common reinforcement learning models like Q-learning (Watkins, Christopher J C H & Dayan, 1992) or the Rescorla-Wagner model (Rescorla & Wagner, 1972). This model architecture implements trial-by-trial prediction errors used to update the hidden variables/internal states of the models carried over discrete time-steps. The prediction errors broadly resemble current neurobiologically plausible computational frameworks proposed in neuroscience like Active Inference (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017) and PC (e.g., Keller & Mrsic-Flogel, 2018). The number of hidden variables and associated computational processes differ between the models proposed in the original study (Apps & Tsakiris, 2013), resulting in different amounts of model parameters (see Tables 5 and 6).

A summary of the computational hypotheses and the associated model-space put forward by the original study can be found in section 2.4.1. This section also translates the psychological hypotheses involving the three proposed factors (view-independence, view-dependence, and context) created in the original study into the associated computational hypotheses. Section 2.4.2 will give details on the implementation of the code and used software packages to ensure the reproducibility of our results. Section 2.4.3 gives details of employed optimization algorithms and model fitting procedure. Section 2.4.4 will discuss the individual model components and respective formulas, i.e., the trial-by-trial updating equations and the overall mathematical behavior of the models. Sections 2.4.5, 2.4.6, and 2.4.7 will discuss the mathematical details of each computational model in more detail. Finally, section 2.4.8 will describe model-selection procedures.

2.4.1. Computational hypotheses

Apps and Tsakiris (2013) proposed a total of six computational models corresponding to six different hypotheses about the computations performed in the brain when the learning of faces occurs. The six models are characterized by different combinations of three main computational processes: view-dependent familiarity learning, view-independent familiarity learning, and contextual learning. View-dependent processing refers to the encoding of face-related visual information as it reaches our retina with little higher-order abstraction, i.e., the same face viewed from different angles induces a new stimulus-representation for each angle and is not recognized as different views of the same stimulus. View-independent processing describes computations that rely more on the abstract higher level of inference in visual perception. Here, it was hypothesized that the brain only uses one abstract representation for face identification, regardless of viewpoint.

In addition to Apps and Tsakiris's (2013) models, we included a model which outputs a 50% answer probability for each of the two answer possibilities at any given trial. These amount to a random-choice model, which will be used as a baseline comparison in all analyses (see Table 5 for an overview of the model space).

2.4.2. Software and implementation

Behavioral data pre-processing and computational models were implemented in Python 3.x on Ubuntu 20.x using various functions from the packages Numpy (Harris et al., 2020), Scipy (Virtanen et al., 2020), Pandas (Jeff Reback et al., 2021), Seaborn (Waskom, 2021) and Pingouin (Vallat, 2018). All code implementations, plots, and raw data for the computational modeling and behavioral analyses can be found in the associated public repository DOI (https://doi.org/10.18112/openneuro.ds004529.v1.1.0), including a full list of dependencies.

2.4.3. Model fitting and optimization

The data was fitted subject-wise using a maximum-likelihood approach. The joint-likelihood of the observations obtained from each subject under each model was evaluated using the sum of the log-likelihoods for the trial-wise model predictions. Parameter optimization for maximizing the joint-probability of the observations under each model was done using the L-BFGS-B algorithm with a finite-differences numerical gradient approximation (including Basin-hopping for improved global optimization) as implemented in the Python package Scipy (Virtanen et al., 2020). Apps and Tsakiris (2013) used a bounded categorical parameter space for each model. Some parameters were varied in their respective space in steps of .01 and others in steps of .1. Due to technical restrictions of our optimization algorithm, the implementation

Table 5 – Model space, computational processes, and the number of associated parameters of each model.

	Hidden variables/computational processes			
	View-dependent processing	View-independent processing	Contextual processing	Number of parameters
View-independent and context-dependent		x	x	4
View-dependent	X			3
View-dependent and context-dependent	X		x	4
View-independent		x		3
View-dependent and view-independent	x	x		5
View-dependent, view-dependent, and context-dependent	x	x	x	6
Random answer probability				0

Table 6 – Individual model parameters. All model parameters are optimized individually/per subject ('idiosyncratic' parameters) and per model if the respective computational process was included in the specific model. See Table 5 for details. Note that some letters are assigned differently than in Apps and Tsakiris (2013).

Computational process	Model parameter					
	Behavioral response	View-dependent face learning		View-independent face learning		Context learning
Description	stochasticity parameter	maximum view-dependent familiarity	view-dependent learning-rate	maximum view-independent familiarity	view-independent learning-rate	context learning-rate
Symbol	β	ρ	γ	λ	α	σ
Bounds	(.1.20)	(.2)	(.1)	(.2)	(.1)	(.1)
Steps in original study	±0.1	±.01	±.01	±.01	±.01	±.01
Steps in our study	±.01	±.01	±.01	±.01	±.01	±.01

of all parameters varied in steps of .01. The bounds of the parameter-space were implemented as closely as described in the Apps and Tsakiris (2013) (see Table 6).

2.4.4. *Model components*

Each model consisted of a hidden variable model and a response model. The architecture of the former differed between the proposed computational models; the latter was identical for all models. The function of the hidden variable model was to update and carry the trial-by-trial estimates of the hidden variables (e.g., contextual familiarity) and to convert them into the total familiarity of a stimulus at a given trial. The response model converted the total familiarity of a stimulus into a model-predicted answer probability for the observed data in this trial. The equations used to calculate contextual familiarity in all computational models were modified by us due to numerical stability issues (see also section '2.3.5 Model 1'). All other equations were unmodified.

2.4.4.1. Response model.

The response model converted the trial-by-trial total stimulus-familiarity predictions outputs by the model into answer probabilities for the two possible actions of the task ('Yes,' 'No'). Apps and Tsakiris (2013) reported their response model as an implementation of the Luce-choice rule (Duncan Luce, 1977; Luce, Ng, Marley, & Aczél, 2008), which in turn is an application of the softmax-function often used in reinforcement learning models (Sutton & Barto, 2018). The model-predicted probability of answering 'Yes' at a given trial was calculated by

$$p(yes|t) = \frac{1}{1 + e^{(-\beta F_t)}} \quad (1)$$

where β is the stochastic parameter indicating the sensitivity of the answer-behavior to the total familiarity of the stimulus at time t (F_t). As implied by the binary answer format, the probability of answering 'No' is given by

$$p(no|t) = 1 - p(yes|t) \quad (2)$$

2.4.4.2. Initialization values of internal model states.

The model architecture used in the original study requires initialization values to start the hidden-variable updating processes. These initialization values may be viewed as the prior information a subject is equipped with (e.g., primed stimulus and context processing) when starting the cognitive task. Contextual familiarity, view-independent familiarity, and view-dependent familiarity all require initialization values. Explicit information was only provided for the view-independent familiarity where the initial hidden variable state $V_{i,0}$ for all stimuli i was defined as

$$V_{i,0} = FPrate^{*}\lambda \quad (3)$$

Here, *FPrate* denotes percentage of false-positive responses for those stimuli presented the first time regardless of view-angle, i.e., wrongly answering that a stimulus was familiar at its initial presentation. λ denotes the individual parameter reflecting the maximum view-independent familiarity achievable for any stimulus in this subject. Analogously, we calculated the initialization values for the view-dependent familiarity by

$$VD_{i,0} = FPrate^{*}\rho \quad (4)$$

where ρ represents the maximum view-dependent familiarity, *FPrate* denotes the view-dependent false-positive answer rate, i.e., the percentage of wrong "yes" answers to the first stimulus presentation while considering different viewing angles as the entire new stimulus representation. The initial contextual familiarity was considered to be 0 due to the total unfamiliarity of each subject with the stimulus material.

2.4.4.3. Perceptual models/hidden variable models.

In this subsection, we will briefly cover the hidden variable models reflecting the computational processes proposed by Apps and Tsakiris (2013). Note that the mathematical notation differs at times from that of the original authors (see Table 7 for detailed information regarding the trial-by-trial model estimates).

2.4.5. *Model 1 - view-independent and context-dependent ('winning-model')*

In model 1, the overall familiarity of a given stimulus is the product of its view-independent familiarity (i.e., perspective invariant face identity learning) and contextual familiarity (i.e., stimulus familiarity on preceding trials). Total familiarity with this model is given by

$$F_t = VI_{i,t}{}^{*}C_t \quad (5)$$

where $VI_{i,t}$ is the view-independent familiarity of stimulus *i* at trial *t*. Note that $VI_{i,t}$ is hypothesized to be calculated after the

Table 7 – Trial-by-trial model estimates of hidden variables. Trial/time is indexed by t. As stimulus familiarities are specific for a given stimulus, index i describes the stimulus-ID in question. All variables are calculated at each trial.

	Hidden variables/Computational processes						
	View-dependent familiarity	View-dependent prediction-error	View-independent familiarity	View-independent prediction-error	Contextual familiarity	Contextual prediction-error	Total familiarity
Model variables	$VD_{i,t}$	τ	$VI_{i,t}$	δ	C_t	ε	F_t

presentation of each stimulus and is subsequently used to calculate the contextual familiarity and, finally, the total familiarity. It is given by the formula

$$VI_{i,t} = VI_{i,n-1} + \alpha\delta \tag{6}$$

where $VI_{i,n-1}$ is the view-independent familiarity of a stimulus i at its previous presentation. α denotes the view-independent learning rate, and δ denotes the view-dependent prediction error. δ is given by

$$\delta = \lambda - VI_{i,n-1} \tag{7}$$

where λ is the individual maximum view-independent familiarity. Contextual familiarity at a given trial t was calculated in the original study via

$$C_t = C_{t-1} + \sigma\varepsilon \tag{8}$$

where C_{t-1} is the contextual familiarity at the previous trial, σ denotes the contextual learning rate, and ε the prediction error. Importantly, the pilot data revealed that formula (8) for the calculation of C_t produced numerically unstable results with trial-by-trial estimates, which approached within a few trials positive or negative infinity (under/overflow errors). We modified the formula to

$$C_t = C_{t-1} - \sigma\varepsilon \tag{9}$$

which yielded numerically stable results. Note that this modified formula was used in our implementation for all models using contextual familiarity estimation. ε, the contextual prediction error, is unmodified and calculated by

$$\varepsilon = C_{t-1} - VI_{i,t} \tag{10}$$

where again C_{t-1} is the contextual familiarity at the last trial and $VI_{i,t}$ is the view-independent familiarity of stimulus i at the current trial calculated by formula (6).

2.4.6. Model 2 - view-dependent

In this model, the total familiarity at a trial was a direct function of the view-dependent familiarity $VD_{i,t}$, where the total stimulus familiarity is calculated by

$$F_t = VD_{i,t} \tag{11}$$

and $VD_{i,t}$ is calculated by

$$VD_{i,t} = VD_{i,n-1} + \gamma\tau \tag{12}$$

where γ denotes the view-dependent learning rate. Similar to the view-independent familiarity, $VD_{i,n-1}$ refers to the view-dependent familiarity of stimulus i at the previous presentation. τ denotes the view-dependent prediction error given by

$$\tau = \rho - VD_{i,t} \tag{13}$$

Here ρ denotes the individual maximum view-dependent familiarity.

2.4.7. Models 3 to 6

All formulas for the other control models (models 3–6) were combinations of the formulas described in the hidden variable models 1 and 2. Therefore, it is unnecessary to describe these models in detail, but only summarize the calculations of the total familiarity (F_t) in Table 8.

2.4.8. Bayesian model selection

We replaced the model selection procedure of Apps and Tsakiris (2013) with an approximate Bayesian model selection via the BIC and an approximate Bayes Factor ABF_{10}. This method will yield intuitive estimates of relative posterior model probability expressed as an approximate Bayes Factor for interpretability. The BIC will be calculated via Formula 14.

$$BIC_M = LL - \frac{1}{2}NP\,ln(N) \tag{14}$$

Here, LL denotes the maximized log-likelihood of the model. NP denotes the number of parameters, and N is the sample size. Note that our BIC definition differs in two aspects from other authors: first, it is multiplied by a factor of 2. Second, we omitted a minus sign. We made these changes so that our BIC can be interpreted directly as an approximate log-evidence for a model (Bishop, 2006).

As mentioned before, we estimated the parameters of the computational models by Apps and Tsakiris (2013) by maximizing the joint probability for the observations under each model. This resulted in a model-specific log-likelihood that will be corrected for model complexity by using the BIC on an individual level. Then, a per-model BIC sum was evaluated, resulting in a vector of group-level but model-specific sums of BIC values, i.e., one sum for each model. Since the Laplace approximation to the model evidence converges asymptotically to the BIC under fairly general conditions, the BIC values can be considered an asymptotic approximation to twice the negative log of the Bayesian posterior probability of a model assuming a uniform model prior (Endres, Chiovetto, & Giese, 2013; Neath & Cavanaugh, 2012). We will denote the BIC for model M as BIC_M for any model M. Here, our main goal using a Bayesian model selection via the BIC was the estimation of the relative posterior probability of a model given all models included in the selection procedure, i.e., $p(M_i|\{M_I\})$ for a model M_i. This posterior probability of a model was calculated by exponentiation of the difference between the model-specific sum of BIC (BIC_M) and the total sum of BIC of all models in the model selection set ($BIC_{\{M_I\}}$). The latter was calculated

Table 8 – Trial-by-trial calculation of total familiarity F_t. Overview of all components used in each model to calculate the total familiarity at a given trial. For descriptions of the individual components refer to Table 7.

Model		Formula
1	View-independent and context-dependent	$F_t = VI_{i,t}{*}C_t$
2	View-dependent	$F_t = VD_{i,t}$
3	View-dependent and context-dependent	$F_t = VD_{i,t}{*}C_t$
4	View-independent	$F_t = VI_{i,t}$
5	View-dependent and view-independent	$F_t = VI_{i,t} + VD_{i,t}$
6	View-dependent, view-independent and context-dependent	$F_t = (VI_{i,t} + VD_{i,t}){*}C_t$

using the LogSumExp function as implemented in the Python package Scipy (Virtanen et al., 2020).

$$P(M_i|\{M_I\}) = exp\left(BIC(M_i|D) - \log\left(\sum_{i}^{I} \exp\left(BIC(M_i|D)\right)\right)\right) \quad (15)$$

For convenience and clear interpretation, this model-specific posterior probability was further transformed into posterior model odds via

$$o(M_i|\{M_I\}) = \frac{p(M_i|\{M_I\})}{1 - p(M_i|\{M_I\})} = ABF_{10} \quad (16)$$

As these odds are in the same range as the BF_{10} , i.e., between 0 and positive infinity, we propose using the conventions of Bayes factors BF_{10} (Lee & Wagenmakers, 2014; adjusted from Jeffreys, 1998). Therefore, the model-specific posterior odds will be called ABF_{10} from here on out. Similar to the BF_{10}, the ABF_{10} can be interpreted as showing the support for our M_{ii}, i.e., that the proposed winning model is most likely given the data and the control models. As per Cortex guidelines, we considered an $ABF_{10} \geq 6$ as a successful replication of the original results. Any value below 6 but above 1 denoted negligible evidence for M_{ii}, and the replication was considered not successful. Any value below 1 denotes evidence for the control models, i.e., that the proposed winning model is not probable given the data and the set of the control models fits the data better.

2.5. MRI data acquisition and analysis

2.5.1. MRI data acquisition

MRI data were acquired with a 32-channel head matrix receive coil on a Siemens 3 Tesla Tim Trio MRI scanner (Siemens Medical Systems, Erlangen, Germany) at the Core-Facility Brain imaging, University of Marburg, Germany. Functional images were collected with a T2*-weighted gradient-echo echo planar imaging (EPI) sequence sensitive to the blood-oxygenation-level-dependent (BOLD) contrast (repetition time (TR) = 1530 ms, echo time (TE) = 30 ms, voxel size = 3 × 3 × 3 mm^3, 48 slices, 3 mm thickness, flip angle = 55°, matrix size = 70 × 70 voxels, field of view (FoV) = 210 × 210 mm^2). Slices were acquired in ascending order parallel to the intercommissural plane (anterior to posterior commissure) using a multiband acceleration factor 2. The initial four images were excluded for both the face identity learning task and the face localizer task from further analyses to remove the influence of T1 stabilization effects. Prior to the functional images, we also acquired a high-resolution T1-weighted anatomical scan covering the whole brain, with a magnetization-prepared rapid gradient-echo (3D MP-RAGE) sequence in the sagittal plane (TR = 1900 ms, TE = 2.26 ms, voxel size = 1 × 1 × 1 mm^3, 176 slices, 1 mm thickness, flip angle = 9°, matrix size = 256 × 256 voxels, FoV = 256 × 256 mm^2). We would like to point out that the imaging parameters for the fMRI sequence differed from those described in stage 1 of the registered report. We decided to choose an imaging sequence that was most appropriate at the time of data collection. The imaging parameters listed in the registered report referred to a now obsolete imaging sequence previously used in our laboratory.

2.5.2. MRI data analysis

2.5.2.1. Image processing. Imaging data were analyzed with SPM12 (www.fil.ion.ucl.ac.uk/spm) within MATLAB (R2017a, https://de.mathworks.com/products/matlab.html) using standard routines and templates. *Face identity learning task*: First, the functional images were slice time corrected and realigned. Second, the high-resolution anatomical image was co-registered with the mean functional image, then segmented and normalized to the MNI standard space. All functional images were normalized using the spatial normalization parameters obtained from the unified segmentation-normalization approach to the anatomical image and resampled to a voxel size of 2 × 2 × 2 mm^3. Finally, all images were smoothed with an isotropic 6 mm full width at half maximum (FWHM) Gaussian kernel. *Functional localizer task*: First, the functional images were realigned. Second, the high-resolution anatomical image was co-registered with the mean functional image, then segmented and normalized to the MNI standard space. All functional images were normalized using the spatial normalization parameters obtained from the unified segmentation-normalization approach to the anatomical image and resampled to a voxel size of 2 × 2 × 2 mm^3. Finally, all images were smoothed with an isotropic 6 mm FWHM Gaussian kernel.

2.5.2.2. Statistical analysis. Statistical analysis of the pre-processed functional data was conducted using a General Linear Model (GLM). In the following, we will describe the statistical analysis separately for the face identity learning task and the face localizer task.

Face identity learning task: The GLM was built to examine the association between the BOLD response and the computational model parameters describing the face learning process best (i.e., 'winning' model, see above).

At the subject level, we included the following conditions in the GLM. The first regressor modeled the onset of the face stimuli. Only those events were included in which a response was given. The second regressor modeled the onset of the

trigger cue events. Again, only those events were included in which a response was given. The third regressor modeled both of the aforementioned events on missed trials. Trials were classified as missed if the participant did not respond to the cue event within a time window of 1000 ms. All three regressors were convolved with the canonical hemodynamic response function implemented in SPM. The fourth regressor describes the effect of specific computational model parameters (e.g., prediction error). This effect—the main effect of interest in this GLM—was implemented as a first-order parametric modulator to the face regressor. Last, we included the six realignment parameters as nuisance regressors to control for movement-related artifacts (regressors 5–10). Furthermore, low-frequency noise in the data was accounted for by a high-pass filter (cut-off frequency: 1/128 Hz). In total, we built—for each computational model—five GLMs, i.e., one GLM for each computational parameter of interest: (1) Overall familiarity of each trial (F_t), (2) view-independent familiarity (V_i), (3) view-independent update (δ), (4) context-dependent familiarity (C_t), and (5) context-dependent prediction error (ε). Additionally, we built another GLM without a parametric modulator. This GLM was used to assess whether the face identify learning task activated face-specific brain regions in the core system (i.e., OFA, FFA, pSTS), as in the face localizer task.

Face localizer task: We used the three regressors (i) faces, (ii) houses, and (iii) scrambled images. Additionally, we included the six realignment parameters as nuisance regressors to control movement-related artifacts. Low-frequency noise in the data was accounted for by a high-pass filter (cut-off frequency: 1/128 Hz). Parameter estimate (ß-) images were computed for each subject. For the face localizer task, we were interested in the conjunction contrast image "faces > houses and faces > scrambled." For these contrasts, we computed t-images for each subject (see Thome et al. (2022) for details). At the group level, the individual contrast images were entered into random-effects analysis using one-sample t-tests.

Next, we calculated the following analyses:

Analysis 1: We used the face localizer to determine the face-sensitive regions OFA, FFA, and pSTS, both for the left and right hemispheres. These regions served as (independently determined) ROIs for the subsequent analyses. In addition, we also determined the Parahippocampal Place Area (PPA) using the contrast "houses > faces." The PPA acted as a control region as its activity should not vary with any of the parameters of our experimental model if activity in the task is driven by familiarity with the identity of the face. All ROIs were determined both on the subject and the group level. For the identification of the ROI center coordinates at the group level, we chose the highest local maximum in specific, anatomically defined regions. The OFA had to be located in the inferior occipital gyrus, the FFA in the fusiform gyrus, the pSTS in the superior posterior sulcus, and the PPA in the parahippocampal gyrus. Anatomical localization of the activated brain regions was assessed using masks created with the WFU-pickatlas (Maldjian et al., 2003, 2004). Precisely, we used for the FFA the fusiform gyrus (IBASPM 116); for the OFA, the inferior occipital gyrus (IBASPM 116); for the STS, the superior and mid-temporal gyrus (IBASPM 116) and for the PPA the parahippocampal gyrus (TD Labels, J. L. Lancaster, Summerlin, Rainey, Freitas, & Fox, 1997; Jack L. Lancaster et al., 2000). For all masks, a 3 d dilation of 2 was applied. For the identification of the ROI center coordinates at the subject level, we chose the subject-specific maximum closest to the respective group maximum, as described in Thome et al. (2022). Again, these maxima had to be in the above-described anatomical masks. These automatically identified coordinates were visually checked for plausibility and, if necessary, corrected (see supplementary material for details). ROIs were then created as spheres with a 10 mm radius around the respective maximum.

Analysis 2: As the main analysis, we assessed whether brain activity at the time of face presentation is modulated by any parameters of the computational model. We expected, analogous to the results of the study of Apps and Tsakiris (2013), that activity in the pSTS varies with the context-dependent familiarity (C_t) (see the introduction, Hypothesis 2), whereas activity in the FFA covaries with the context-dependent prediction error (ε) (Hypothesis 3). Group analyses were calculated in two different ways. In the first analysis, we calculated for each single-subject ROI the mean ß-parameter of interest (i.e., averaged over all voxels in the respective ROI). Using a one-sample t-test, we then assessed whether these parameters significantly differed from zero. This approach is less flexible than the standard voxel-wise analysis (described below) since it does not allow the assessment of brain activity outside the ROIs. It is, however, potentially more sensitive since it takes into account individual differences in the anatomical location of the face-sensitive regions. As a statistical threshold, we set $p < .05$, corrected for the number of statistical tests. In the second analysis, we used the standard voxel-wise approach using the second-level fMRI models. In the first step, activity was assessed in the above-defined group-level ROIs (using a threshold of $p < .05$, corrected for multiple comparisons at the voxel level). In a second step, we also assessed BOLD activity at the whole-brain level (using a threshold of $p < .001$ uncorrected).

2.5.2.3. Assessment of MRI data quality. *MRI scanner quality assurance*: MRI scanner stability was assessed by a general, non-study-specific quality assurance protocol. This protocol involved the regular measurements of both an ACR phantom and a gel phantom at fixed time points. Quality metrics (e.g., spatial signal-to-noise ratio, temporal signal-to-noise ratio) were calculated using the *LABQA2GO* toolbox (see Vogelbacher et al., 2018; Vogelbacher et al., 2019 for details).

General MRI data quality assurance: Routinely, all study-specific MRI data were visually inspected for artifacts and altered morphology, potentially affecting further analyses. To complement our visual inspection, the MRIQC tool (Esteban et al., 2017) was used to extract image quality metrics (IQMs) and visual reports that would allow us objective quality control of the neuroimaging data.

Study-specific MRI data quality assurance: In the present study, we further used neutral outcome measures to ensure good data quality, in particular for the face identity learning experiment. These outcome measures comprised the behavioral learning effect, the results of the localizer task, and the "face contrast" from the face identity learning experiment. We

expected participants to show (1) a behavioral learning effect above chance. This is defined when a subject reaches a level of correct face recognition of at least 60% averaged over all stimuli at the end of the experiment. More specifically, the percentage of correct responses was averaged for the last three presentations of each stimulus for stimuli that are presented more than once. This stimulus-specific average was then, in turn, averaged again over all relevant stimuli resulting in a single average percentage indicating the overall highest mean recognition performance reached by a subject at the end of the experiment (see 2.3.1, behavioral data quality, for details). We further expected (2) BOLD activity in the core system in the face localizer experiment and (3) BOLD activity in the core system in the face identity learning experiment for the contrast "faces > baseline." In both activation patterns, we determined whether brain activity can be found in the left OFA, right OFA, left FFA, right FFA, left pSTS, and right pSTS. The brain activation patterns were first set to a conservative threshold of $p < .05$, corrected for multiple comparisons at the whole-brain level (family-wise error, FWE, corrected at the voxel level). If BOLD activity was not found in all regions of the core system at this threshold, the p-value was subsequently lowered to more liberal thresholds ($p < .001$ and $p < .05$, respectively, uncorrected for multiple comparisons) (see Schuster et al., 2017; Hildesheim et al., 2020 for an extensive discussion of this procedure). Data from sessions in which participants failed to fulfill these criteria were excluded.

We excluded three subjects from the final data analysis. One subject (ID 21) was excluded because of a face recognition rate below 60%. Two participants (ID 16 and ID 26) were excluded because of low fMRI data quality (framewise displacement >.3, cf. MRIQC results in online repository at: https://osf.io/tye24/). Three subjects (ID 4, ID 18, and ID 28) were further excluded from the fMRI data analysis. When plotting the individual regression weights for the parametric modulators, we encountered that these subjects had absurdly high values. Upon closer inspection, we found out that in all cases, one of the trial-by-trial parameters extracted from the fitted winning model (view-independent familiarity , $VI_{i,t}$) showed little to no variation over trials. This most likely led to convergence issues when estimating covariation with BOLD activity.

3. Results

In the following, we will present behavioral results (3.1), computational modeling results (3.2), and fMRI results (3.3). We will first give the results of the primary analyses outlined in the registered report in each section. Exploratory analyses will be clearly distinguished from this and are presented in a second step. Importantly, the data and analysis scripts of this study are publicly available at: https://osf.io/tye24/.

3.1. *Behavioral results*

In the final sample, the subjects showed an apparent learning effect in the face-learning task (Fig. 5 left). With each subsequent stimulus presentation, the tendency to recognize a face increased, exceeding chance-level answering after only two presentations (see Fig. 5).

When focusing on the percentage of answering 'yes' at the first three stimulus presentations, i.e., 1st,2nd,3rd ($M = .489$, $SD = .107$) versus the last three stimulus presentations, i.e., 10th, 11th, 12th ($M = .842$, $SD = .106$) this difference is significant and meaningful indicating a fast learning effect. Specifically, a paired sample two-sided t-test revealed a significant increase in the tendency to give 'yes' responses with $t(30) = 15.697$, $BF_{10} > = 100$, $p = 5.204430 \times 10^{-16}$, $d = 3.297$ from the first three presentations vs. the last three (see Table 9).

This effect could not be explained by subjects responding habitually 'yes' more often as the experiment progressed. The proportions of novel faces presented throughout the task and, consequently, the ground truth familiarity varied (cf. Fig. 3) over the course of the task. However, we additionally performed a two-sided paired Wilcoxon signed-rank test on the probability of answering 'yes' (i.e., reporting to have recognized a face) between the first half and second half of the task

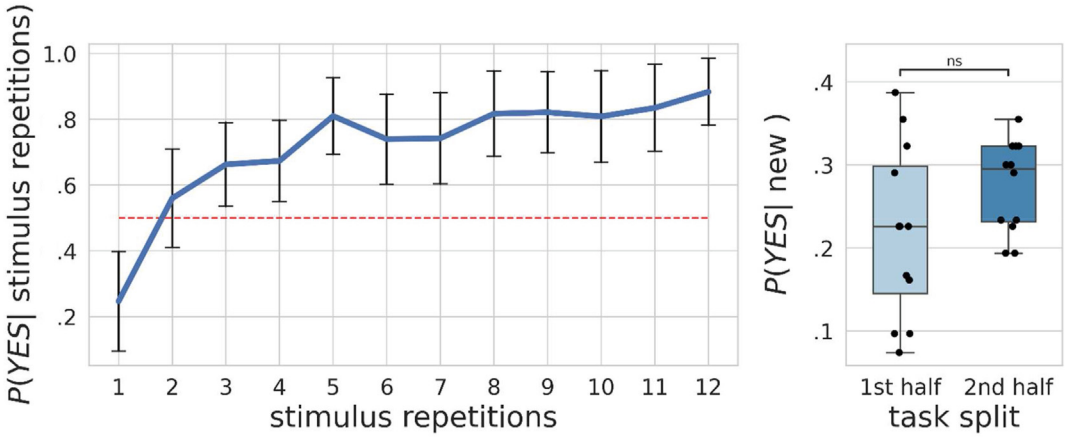

Fig. 5 – Behavioral results from face identity learning task. Left plot: Group-level mean probability of face recognition (answering 'yes') given the number of stimulus repetitions. Error bars are standard deviations. The Red dashed line marks chance-level responding, i.e., 50%. Right plot: Group-level mean probability of face recognition (answering 'yes') given no prior stimulus repetitions split for the first and second half of the task.

Table 9 – Group-level task performance in the face identity learning task. Falsely remembered trials describe the erroneous response that the face is known at its first presentation. Not remembered trials describe the case where a face id was shown at least once and not recognized. Missed answers represent the number of trials where the subject did not provide any answer.

% correct answers		No. of total errors		No. falsely remembered stimuli		No. of not remembered stimuli		No. of missed answers	
M	SD	M	SD	M	SD	M	SD	M	SD
.758	.070	58.483	15.924	5.870	3.612	39.516	14.009	1.129	1.565

exclusively for new stimuli. The test confirmed that the probability of reporting to have recognized a face did not significantly differ between the first half (MDN = .226) and second half (MDN = .295) half with $W = 18$, $p = .11$, and a non-parametric common language effect size (McGraw & Wong, 1992) of CLES = .32.

3.2. *Computational modeling*

With our computational modeling approach, we found positive evidence for Hypothesis 1 of our replication. We defined a successful replication as an $ABF_{10} \geq 6$ (highest approximate Bayes factor, see section 2.4.8 for definition) for the winning model of the original study (i.e., the view-independent and context-dependent model) on a group level. Indeed, we found an $ABF_{10} = 9999.50$, which corresponds to a posterior model probability of >99.999%, given the observed data and other models in the approximate Bayesian model selection procedure. To avoid floating point underflows in the calculation, we capped the ABF results to a maximum value of 9999.50.

We performed the primary approximate Bayesian model selection described in the methods section 2.4.8. and another exploratory model selection, excluding the random-choice model, as it was not included by the original authors. However, excluding this additional baseline model did not influence the conclusions from the confirmatory model selection results.

On another exploratory note, a more heterogeneous picture emerges when focusing on subject-level results. Even though the view-independent and context-dependent model emerged as having the highest ABF_{10} in our model selection, this model only fitted 12 of the 31 subjects included in the final sample best. We determined this by selecting the model with the largest BIC (i.e. closest to zero) on a subject level following the winner-takes-all-principle. The control model, including view-dependent and contextual processing, fitted 13 of the 31 final subjects best. Additionally, the control model encoding only view-independent without contextual modulation best fitted four out of the subjects. For two subjects, none of the models provided a better model fit than the random choice model.

In summary, we were able to corroborate Hypothesis 1 at the group level and the main computational modeling results obtained by Apps and Tsakiris (2013) following our predefined cut-off for successful replication of $ABF_{10} \geq 6$. Additionally, following the $BF_{10} \geq 150$ convention by Kass and Raftery (1995), our evidence can be considered *'very strong'* on a group level.

3.3. *fMRI results*

In the following, we present the results from the face localizer task (3.3.1), then the face identity learning task (3.3.2).

3.3.1. *Face localizer*

Based on independent data, the face localizer task was included in the experiment to determine face-sensitive ROIs for analyzing the face identity learning task. It was associated with BOLD activity in a distributed network encompassing the bilateral occipito-temporal cortex (including the core system's brain regions OFA, FFA, and pSTS) and frontal and parietal areas (see Fig. 7).

The localization of the OFA, FFA and pSTS was determined for the left and right hemispheres both at the group level and the individual subject level using the conjunction contrast "face > houses" AND "faces > scrambled." The individual center coordinates are depicted for each ROI in Fig. 6.

In Table 10, we present the center coordinates of each ROI at the group level of the face-localizer task. The

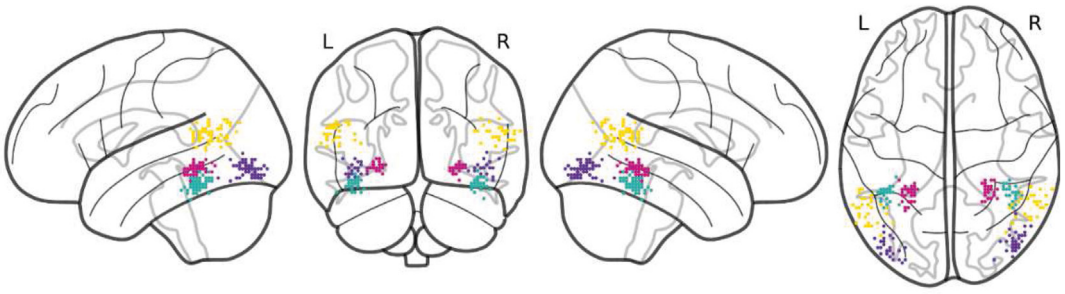

Fig. 6 – Individual maxima for the three brain regions of the core system of face perception, as assessed by the face localizer task: OFA (purple), FFA (green), STS (yellow), and PPA (pink). Coordinates were visualized with nilearn (version .7.0) (Abraham et al., 2014)

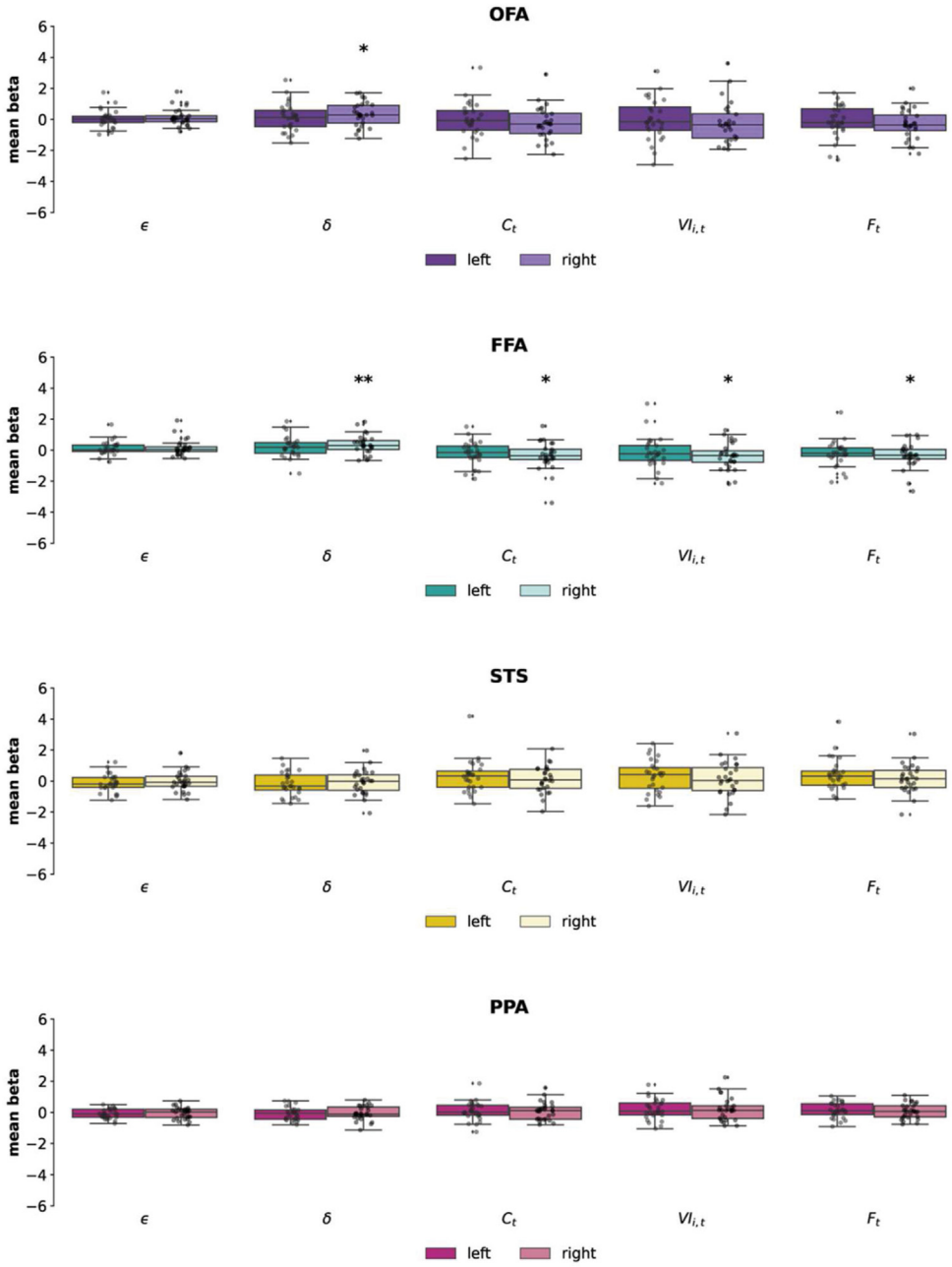

Fig. 7 – In our main analysis, we investigated whether face-related brain activity was modulated by trial-by-trial estimates of the winning computational model i.e. the model involving view-independent and contextual processing. Here, we present the mean regression weights ("mean beta") for each ROI and each parameter. Using one sample *t*-tests, we assessed which modulation parameters were significantly different from zero (*: $p < .05$, **: $p < .01$).

supplementary material/ Appendix G also presents each ROI's center coordinates for each subject. In addition, we also determined the Parahippocampal Place Area (PPA) using the contrast "houses > faces." The PPA acts as a control region, as its BOLD activity should not vary with any of the parameters of our experimental model.

3.3.2. *Face identity learning task*

The face identity learning task was, just like the face localizer task, associated with BOLD activity in a distributed network encompassing the core system's brain regions (contrast "faces > baseline"; see Table 11 for a list of coordinates at the group level, supplementary material/ Appendix E for an illustration of the activation pattern). The anatomical localization of the regions of the core system was, as expected, highly similar when determined by the face localizer task and the face identity learning task.

In our main analysis, we assessed whether the face-related BOLD activity is modulated by parameters of the winning computational model, i.e., the model involving view-

Table 10 – Group analysis. For the core system, the conjunction contrast ("faces > houses" AND "faces > scrambled") was used. For the PPA, the contrast "houses > faces" was used. Bilateral FFA, pSTS, and PPA were detectable at $p < .05$, FWE corrected at the voxel level. The left OFA was only found at $p < .05$ uncorrected, and the right OFA at $p < .001$ uncorrected.

Face localizer task: MNI coordinates of BOLD activity in the core system (OFA, FFA, pSTS) and the PPA				
Region of interest	x	y	z	*t-value*
Left OFA	−42	−84	−10	1.82
Right OFA	48	−74	−6	3.86
Left FFA	−42	−46	−18	7.32
Right FFA	42	−46	−16	6.80
Left pSTS	−58	−54	8	5.55
Right pSTS	46	−54	14	6.25
Left PPA	−22	−46	−8	17.57
Right PPA	24	−46	−10	16.95

independent and contextual processing. Our main hypotheses were that activity in the right pSTS covaries with the trial-by-trial model estimate of contextual familiarity (C_t) (for mathematical description, see methods section, **Hypothesis 2**), whereas activity in the FFA covaries with the view-independent prediction error (δ) (**Hypothesis 3**).

In the first step, we calculated, for each subject and each ROI, the beta regression weight of the GLM for the respective computational model parameter averaged across all voxels in the ROI. ROIs were created as spheres with a 10 mm radius around the individual center coordinates determined from the face localizer task. Using one-sample *t*-tests, we assessed whether these parameters were at the group level significantly different from zero. If so, it would indicate that these specific model parameters are significantly associated with the BOLD activity. The results are presented in Fig. 7. A summary of statistics (e.g., mean, standard deviation, *p*-values) can be found in the supplementary material/ Appendix E and F. The main findings can be summarized as follows: **First**, we did not find a significant association between activity in the right pSTS and context-dependent familiarity (C_t). **Hypothesis 2** was therefore not confirmed. We also did not find a significant association between activity in the right or left pSTS with any other computational model parameter. **Second**, we found that activity in the right FFA covaried with the view-independent prediction error (δ). **Hypothesis 3** was therefore confirmed at least for the right FFA, albeit not for the left. BOLD activity in the right FFA further covaried with δ, $VI_{i,t}$ and F_t. **Third**, we further found that activity in the right OFA was associated with δ. Activity in the PPA, a non-face-sensitive region, did not significantly covary with any parameter.

In a second step following the main authors in a more exploratory path, we additionally applied a standard voxel-wise approach using second-level fMRI models (i.e., one-sample *t*-tests for each computational parameter). This approach is also able to assess associations between computational model parameters and BOLD activity outside the core system's brain regions. We decided to specifically present results for the trial-by-trial estimates δ and C_t only from the winning model since our main hypotheses were based on these parameters. We used a statistical threshold of $p < .05$ family-wise error rate corrected for multiple comparisons at the voxel level, either at the whole brain level or with a small volume correction (svc); for svc, we used, in accordance with Apps and Tsakiris (2013), a mask consisting of voxels being activated at $p < .001$ for the conjunction contrast "faces > houses" AND "faces > scrambled" derived from the face localizer task. This ensured that only voxels were analyzed that could be considered face-sensitive. For C_t no cluster survived this threshold. For δ, we found significant brain activation in the left superior frontal gyrus, left caudate, right amygdala, and right FFA (Table 12). In the supplementary material, we additionally present activated clusters at an uncorrected threshold of $p < .001$ (Fig. 7 and supplementary material/ Appendix F).

4. Discussion

Apps and Tsakiris (2013) investigated the computational and neural mechanisms underlying the learning of new face identities. On the one hand, they showed that behavioral responses on a face recognition task could be predicted by the level of contextual and facial familiarity in a computational model derived from predictive coding principles. On the other hand, they showed that variations in key parameters of this model (e.g., prediction error) were associated with BOLD

Table 11 – Group analysis, GLM without parametric modulator, contrast "faces > baseline." All clusters but the left pSTS were detectable at $p < .05$ FWE corrected at the voxel level. The left pSTS was only found at $p < .001$ uncorrected.

Face identity learning task: MNI coordinates of BOLD activity in the core system (OFA, FFA, STS)				
Region of interest	x	Y	z	*t-value*
Left OFA	−40	−82	−8	20.21
Right OFA	42	−78	−6	16.36
Left FFA	−36	−44	−20	14.74
Right FFA	36	−40	−16	15.04
Left pSTS	−52	−46	14	4.73
Right pSTS	46	−40	14	10.27

Table 12 – T-test group analysis for the parametric modulator δ. Significant clusters at $p < .05$ FWE corrected at the voxel level are marked with ** and significant clusters after small volume correction with the conjunction contrast (faces > houses AND faces > scrambled) derived from the localizer task ($p < .001$ uncorrected) are marked with *.

Group activation of view independent familiarity (δ) identity learning task, MNI (x, y, z)				
Region	x	y	z	*t-value*
Left superior frontal gyrus medial segment	−6	34	44	6.44**
Left caudate	−20	8	4	6.56**
Right amygdala	20	−4	−14	6.65**
Right FFA	42	−50	−16	4.99*

activity changes in specific areas of the face perception network. The present preregistered study aimed to replicate these findings. In the following, we will first present the results we could replicate and those we could not (4.1). We will then shortly outline which new insights we gained about the face perception network and describe open research questions (4.2).

4.1. A comparison of the results of Apps and Tsakiris (2013) with the present replication study

In their study, Apps and Tsakiris (2013) presented subjects with previously unknown faces. They were instructed to remember the faces. Some stimuli were repeated throughout the experiment so that they became increasingly familiar. For each face, subjects had to respond whether or not they had previously seen this face during the experiment. The authors systematically varied contextual familiarity by the general level of stimulus familiarity in the immediate history of a given trial (i.e., the number of more familiar faces previously presented) and stimulus familiarity by the number of times a specific face identity had been shown before. The main results reported by Apps and Tsakiris (2013) that we set out to replicate were the following: **First**, behavioral responses on a face recognition task could be modeled best by a behavioral computational model assuming a multiplicative effect of contextual familiarity (C_t) and facial familiarity irrespective of the viewing angle ($VI_{i,t}$). **Second**, activity in the right STS was correlated with that model's trial-by-trial contextual familiarity parameter (C_t). **Third**, activity in both the right and the left FFA was correlated with the trial-by-trial prediction error estimate updating facial familiarity (δ).

Hypothesis 1: Computational modeling of behavioral results. Similarly, to the original study, we evaluated which of the proposed computational models explained the learning processes best and most parsimonious. On the group level, we could replicate the original results by Apps and Tsakiris (2013). The model involving abstracted and perspective-invariant facial representations (view-independent processing) and contextual influences captured the learning process best on a group-level. The contextual influence was operationalized as the level of familiarity of the preceding facial identities, which is hypothesized to provide additional information for an agent when predicting the familiarity of the upcoming facial stimulus. Notably, the winning model is consistent with fundamental concepts of the (non-hierarchical) predictive coding framework (e.g., prediction errors). Supplementary analyses further highlighted that the trial-by-trial model parameters extracted from the winning model covaried with the observed behavior. In summary, we were able to replicate Hypothesis 1 decisively.

Hypotheses 2 and 3: Association between model parameters and BOLD activity. Whether the results of an experiment can be replicated also depends on the data quality of the replication study. A non-replication might happen, of course, simply because the acquired data is of insufficient quality. However, we have no evidence that this study's data quality was poor, as assessed, for instance, by the MRIQC analyses. The activation patterns of the basic contrasts ("faces > baseline") also provided highly plausible results. The face localizer task (which was used to determine the ROIs) was associated with robust activity in the core system of face perception. It was possible to determine the core regions OFA, FFA, and pSTS, as well as the control region PPA in both hemispheres for each subject. The presentation of faces was associated with activation in the face perception network, particularly in the core system. Therefore, there was no evidence that a possible non-replication of previous results was due to poor data quality.

Hypothesis 2 (Activity in the STS): Apps and Tsakiris (2013) showed that activity in the right STS covaried with the trial-by-trial model estimate of contextual familiarity (C_t). We did not replicate this finding ($p = .58$, see supplementary material/ Appendix E and F). Exploratory analyses further showed that activity in the right pSTS also did not covary with any other model parameter. An exploratory whole-brain analysis showed that contextual familiarity (C_t) was not significantly correlated with activity in any other brain region outside the core system of face perception. In summary, we were not able to replicate Hypothesis 2.

To further understand the discrepancies between both studies, we analyzed whether there were differences in the localization of the face-sensitive regions in the STS. Although these regions were determined in both studies using a similar procedure, their localization differed. In the present study, we found activity in the posterior part of the STS, as typically reported in studies on face perception (Fox, Iaria, & Barton, 2009; Haxby et al., 2000). In contrast, Apps and Tsakiris (2013) located the STS in far more anterior regions. It might therefore be possible that our study did not replicate the previous results simply because we analyzed activity in a different region. However, as mentioned before, both regions were defined using the same approach. Using the whole brain analysis approach, we further analyzed the association between contextual familiarity and BOLD activity. Nevertheless, at an uncorrected threshold of $p = .05$, we did not find any activity in any part of the STS.

We have also to acknowledge, however, that the actual power for the STS was lower than for the FFA. The expected effect size (i.e., Cohen's d) was averaged across all three ROIs reported in the original study (i.e., right STS, left FFA, right FFA), yielding $d = .788$. When considering the STS separately, the expected effect size would have been .403, leading to an actual power for the STS of .37. Our study thus had effectively less power to find the STS effect compared to the FFA effect.

Hypothesis 3 (Activity in the FFA): Apps and Tsakiris (2013) showed that activity in both the right and the left FFA was correlated with the trial-by-trial view-independent prediction error estimate updating facial familiarity (δ). We also found that activity in the right FFA significantly covaried with this parameter (δ), thus confirming Hypothesis 3, at least for the right FFA. We did, however, not find a significant effect for the left FFA as previously reported ($p = .11$, see supplementary material/ Appendix E and F). Exploratory analyses further showed that BOLD activity in the right FFA also covaried with

other model parameters (i.e., with δ, $VI_{i,t}$ and F_t). The contextual prediction error ε was the only computational parameter for which we did not find an association ($p = .15$, see supplementary material/ Appendix F). It has to be noted, however, that the computational parameters partly covaried across trials (see supplementary material/ Appendix I). Therefore, it is not possible to conclude that each parameter, even if significant, reflects a unique contribution from that parameter. The right FFA might only encode, for instance, facial familiarity, which happens to be correlated with other model parameters.

Furthermore, an exploratory whole-brain analysis showed that the association between the view-independent prediction error estimate (δ) and BOLD activity was not restricted to the right FFA but was also found in a large network outside the core system of face perception. This network encompasses brain regions in the left superior frontal gyrus, left caudate, and right amygdala. In summary, we could replicate Hypothesis 3 for the right FFA but not for the left FFA.

4.2. *New insights into the face perception network and open research questions*

The recognition of familiar faces is crucial for appropriate social interactions. The computational processes that underlie the learning of facial identities, i.e., the transition of faces from unfamiliar to familiar, are still poorly understood in humans. Formally, these processes can be modeled within the meta-framework of the Hierarchical Mechanistic Mind (Badcock, Friston, & Ramstead, 2019; Badcock, Friston, Ramstead, Ploeger, & Hohwy, 2019) encompassing the recent neuro-theoretic frameworks of the Bayesian brain, predictive coding, and active inference under the free-energy principle (see the introduction for a thorough overview). The empirical foundation for PC consistency in face processing, however, is sparse. A notable exception is the study by Apps and Tsakiris (2013). The authors used, for the first time, a combination of computational modeling and neuroimaging to characterize, both at the behavioral and at the brain level, fundamental computational mechanisms underlying the perceptual learning of faces.

In the long term, we want to use similar approaches to understand better mental disorders (e.g., autism spectrum disorder), leveraging theory-driven computational psychiatry methods (i.e., translational neuromodeling approaches) following process theories based on the HMM framework (Haker, Schneebeli, & Stephan, 2016). Before embarking on these projects, we wanted to improve the empirical database and ultimately ensure that we worked with paradigms that would yield replicable results. More specifically, we wanted to test whether crucial model parameters of computational models describing face perception processes are associated with specific brain activity patterns.

Our study supports the original findings of Apps and Tsakiris (2013) in two critical aspects. We showed that behavioral responses in the face recognition task could be predicted by the level of contextual and facial familiarity in a computational model derived from predictive coding principles. On the other hand, we showed that trial-by-trial variations of these model parameter estimates covaried with BOLD activity in specific, face-sensitive brain regions. Additionally, we presented evidence that face identity learning likely follows predictive coding principles on a computational level and that crucial parameters of this computational model can be localized to specific brain regions.

Our study, however, paints a different picture of how these computational model parameters are reflected in the brain. Apps and Tsakiris (2013) suggested that a particular parameter is associated with a particular brain region at a time. More specifically, they associated the STS with contextual familiarity and the FFA with the prediction error updating facial familiarity ("double dissociation of brain functions"). They concluded that "an important functional property of the FFA is to update how familiar a facial identity is […] while the STS processes task-relevant, contextual information about the recent history of the familiarity of faces" (Apps & Tsakiris, 2013, p. 7). In contrast, our study comes to a different conclusion. Our results show that BOLD activity in the right FFA is associated with multiple computational model parameters, while we do not find an association with STS activity.[3] Our data do not support any dissociation between particular brain regions in the sense that one region is specific for one aspect, another region for another aspect of the model.

In future work, we aim to investigate further aspects of the compatibility of predictive processes with facial identity learning. **First**, we will test whether other computational modeling approaches consistent with hierarchical predictive coding (e.g., hierarchical Gaussian filter, C. Mathys, Daunizeau, Friston, & Stephan, 2011; C. D. Mathys et al., 2014) are better suited to model behavioral data. If so, we plan to assess how these model parameters are related to the BOLD activity of the face-processing areas. **Second**, we will further investigate whether interindividual differences in the winning models are also associated with differences on the neural system level. As a reminder, even though the view-independent and context-dependent model emerged as having the highest posterior weight in our model selection, it only fitted 12 of the 31 subjects included in the final sample best.

Open Practices Section

The study in this article earned Open Data, Open Material and Preregistered badges for transparent practices. The data and materials used in this study are available at: https://osf.io/tye24/ and Preregistration is available under the link: https://doi.org/10.17605/OSF.IO/A8VU7

Funding & acknowledgements

The research project is funded by German Research Foundation (DFG), grant number 417284407.

[3] To avoid misinterpretations of our findings, we would like to note that the computational parameters partly covary across trials and therefore one cannot conclude that single-parameter based significances reflect a unique contribution of that parameter. See also Appendix I.

Contributor role taxonomy (CRediT)

Nestor Zaragoza-Jimenez: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, Data Curation **Hauke Niehaus:** Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, Data Curation **Ina Thome:** Methodology, Software, Formal analysis, Writing - Original Draft, Data Curation **Christoph Vogelbacher:** Writing - Review & Editing, Supervision **Gabriele Ende:** Writing - Review & Editing, Supervision, Funding acquisition **Inge Kamp-Becker:** Writing - Review & Editing, Supervision, Funding acquisition **Dominik Endres:** Conceptualization, Supervision, Methodology, Funding acquisition, Writing - Review & Editing **Andreas Jansen:** Conceptualization, Formal analysis, Supervision, Methodology, Funding acquisition, Data Curation, Writing - Original Draft, Writing - Review & Editing.

Supplementary data

Supplementary data and appendices to this article can be found online at https://doi.org/10.1016/j.cortex.2023.05.021.

REFERENCES

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics, 8*, 14. https://doi.org/10.3389/fninf.2014.00014

Apps, M. A., & Tsakiris, M. (2013). Predictive codes of familiarity and context during the perceptual learning of facial identities. *Nature Communications, 4*. https://doi.org/10.1038/ncomms3698

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews, 31*, 104–121. https://doi.org/10.1016/j.plrev.2018.10.002

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective & Behavioral Neuroscience, 19*(6), 1319–1351. https://doi.org/10.3758/s13415-019-00721-3

Barron, H. C., Auksztulewicz, R., & Friston, K. (2020). Prediction and memory: A predictive coding account. *Progress in Neurobiology, 192*, Article 101821. https://doi.org/10.1016/j.pneurobio.2020.101821

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. https://cds.cern.ch/record/998831.

Brodski-Guerniero, A., Paasch, G.-F., Wollstadt, P., Özdemir, I., Lizier, J. T., & Wibral, M. (2017). Information-theoretic evidence for predictive coding in the face-processing system. *The Journal of Neuroscience: the Official Journal of the Society for Neuroscience, 37*(34), 8273–8283. https://doi.org/10.1523/JNEUROSCI.0614-17.2017

Burns, E. J., Arnold, T., & Bukach, C. M. (2019). P-curving the fusiform face area: Meta-analyses support the expertise hypothesis. *Neuroscience and Biobehavioral Reviews, 104*, 209–221. https://doi.org/10.1016/j.neubiorev.2019.07.003

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annu Rev Vis Sci, 1*, 393–416. https://doi.org/10.1146/annurev-vision-082114-035518

Duncan Luce, R. (1977). *The choice axiom after twenty years*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.336.1366&rep=rep1&type=pdf.

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience, 30*(49), 16601–16608. https://doi.org/10.1523/JNEUROSCI.2770-10.2010

Endres, D. M., Chiovetto, E., & Giese, M. A. (2013). Model selection for the extraction of movement primitives. *Frontiers in Computational Neuroscience, 7*, 185. https://doi.org/10.3389/fncom.2013.00185

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). Mriqc: Advancing the automatic prediction of image quality in MRI from unseen sites. *Plos One, 12*(9), Article e0184661. https://doi.org/10.1371/journal.pone.0184661

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. *Human Brain Mapping, 30*(5), 1637–1651. https://doi.org/10.1002/hbm.20630

Frank, D., & Kafkas, A. (2021). Expectation-driven novelty effects in episodic memory. *Neurobiology of Learning and Memory, 183*, Article 107466. https://doi.org/10.1016/j.nlm.2021.107466

Freiwald, W. A. (2020). The neural mechanisms of face processing: Cells, areas, networks, and models. *Current Opinion in Neurobiology, 60*, 184–191. https://doi.org/10.1016/j.conb.2019.12.007

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation, 29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Gluth, S., & Jarecki, J. B. (2019). On the importance of power analyses for cognitive modeling. *Computational Brain & Behavior, 2*(3), 266–270. https://doi.org/10.1007/s42113-019-00039-w

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience, 19*(11), 1803–1814. https://doi.org/10.1162/jocn.2007.19.11.1803

Haker, H., Schneebeli, M., & Stephan, K. E. (2016). Can bayesian theories of autism spectrum disorder help improve clinical practice? *Frontiers in Psychiatry, 7*, 107. https://doi.org/10.3389/fpsyt.2016.00107

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Haxby, J. V.[J. V.], Hoffman, E. A., & Gobbini, M. I.[M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science, 4*(6), 223–233. https://doi.org/10.1016/s1364-6613(00)01482-0

Hildesheim, F. E., Debus, I., Kessler, R., Thome, I., Zimmermann, K. M., Steinsträter, O., et al. (2020). The trajectory of hemispheric lateralization in the core system of face processing: A cross-sectional functional magnetic resonance imaging pilot study. *Frontiers in Psychology, 11*,

Article 507199. https://doi.org/10.3389/fpsyg.2020.507199
Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 35*(2), 209–223. https://doi.org/10.1111/mila.12281
Isager, P. M., van 't Veer, A. E., & Lakens, D. (2021a). *Replication value as a function of citation impact and sample size*. https://doi.org/10.31222/osf.io/knjea
Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., et al. (2021b). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000438
Issa, E. B., Cadieu, C. F., & DiCarlo, J. J. (2018). Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *ELife, 7*. https://doi.org/10.7554/eLife.42870
Jeffreys, H. (1998). *The theory of probability*. Oxford: OUP.
Kaliukhovich, D. A., & Vogels, R. (2011). Stimulus repetition probability does not affect repetition suppression in macaque inferior temporal cortex. *Cerebral Cortex, 21*(7), 1547–1558. https://doi.org/10.1093/cercor/bhq207
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997
Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. Lond. B Biol. Sci., 361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *The Journal of the Acoustical Society of America, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572
Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron, 100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003
Lancaster, J. L.[J. L.], Summerlin, J. L., Rainey, L., Freitas, C. S.[C. S.], & Fox, P. T.[P. T. (1997). The Talairach Daemon a database server for talairach atlas labels. *Neuroimage, 5*(4), S633. PART II).
Lancaster, J. L.[J. L.], Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S.[C. S.], Rainey, L., et al. (2000). Automated Talairach Atlas labels for functional brain mapping. *Human Brain Mapping, 10*(3), 120–131. https://doi.org/10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8
Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759
Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 68*, 155–168. https://doi.org/10.1016/j.cortex.2015.02.014
Luce, R. D., Ng, C. T., Marley, A. A. J., & Aczél, J. (2008). Utility of gambling I: Entropy modified linear weighted utility. *Economical Theory, 36*(1), 1–33. https://doi.org/10.1007/s00199-007-0260-5
Maldjian, J. A., Laurienti, P. J., & Burdette, J. H. (2004). Precentral gyrus discrepancy in electronic versions of the Talairach atlas. *Neuroimage, 21*(1), 450–455. https://doi.org/10.1016/j.neuroimage.2003.09.032
Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage, 19*(3), 1233–1239. https://doi.org/10.1016/S1053-8119(03)00169-1
Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E.[K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5*, 39. https://doi.org/10.3389/fnhum.2011.00039
Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., et al. (2014). Uncertainty in perception and the hierarchical Gaussian filter. *Frontiers in Human Neuroscience, 8*, 825. https://doi.org/10.3389/fnhum.2014.00825
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361
Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 108*(48), 19401–19406. https://doi.org/10.1073/pnas.1112895108
Neath, A. A., & Cavanaugh, J. E. (2012). The bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics, 4*(2), 199–203. https://doi.org/10.1002/wics.199
O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences, 6*. https://doi.org/10.1016/s1364-6613(02)01908-3
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4
Pitcher, D., Walsh, V., & Duchaine, B.[B.] (2011). The role of the occipital face area in the cortical face perception network. *Experimental Brain Research, 209*(4), 481–493. https://doi.org/10.1007/s00221-011-2579-1
Rauss, K., Schwartz, S., & Pourtois, G. (2011). Top-down effects on early visual processing in humans: A predictive coding framework. *Neuroscience and Biobehavioral Reviews, 35*(5), 1237–1253. https://doi.org/10.1016/j.neubiorev.2010.12.011
Reback, J., jbrockmendel, W. M. K., Joris Van den Bossche, Augspurger, T., Cloud, P., Hawkins, S., et al. (2021). *pandas-dev/pandas:* Pandas 1.3.4 *[Computer software]*. Zenodo.
Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement*. https://www.researchgate.net/publication/239030972_A_theory_of_Pavlovian_conditioning_The_effectiveness_of_reinforcement_and_non-reinforcement.
Said, C. P.[C. P.], Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V.[J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *J. Vis., 10*(5), 11. https://jov.arvojournals.org/article.aspx?articleid=2121013.
Schuster, V., Herholz, P., Zimmermann, K. M., Westermann, S., Frässle, S., & Jansen, A. (2017). Comparison of fMRI paradigms assessing visuospatial processing: Robustness and reproducibility. *Plos One, 12*(10), Article e0186344. https://doi.org/10.1371/journal.pone.0186344
Schwiedrzik, C. M., & Freiwald, W. A. (2017). High-level prediction signals in a low-level area of the macaque face-processing hierarchy. *Neuron, 96*(1), 89–97.e4. https://doi.org/10.1016/j.neuron.2017.09.007
Shipp, S., Adams, R. A., & Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neurosciences, 36*(12), 706–716. https://doi.org/10.1016/j.tins.2013.09.004
Summerfield, C., & Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron, 59*(2), 336–347. https://doi.org/10.1016/j.neuron.2008.05.021
Summerfield, C., & Lange, F. P. de (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews. Neuroscience, 15*(11), 745–756. https://doi.org/10.1038/nrn3838

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience, 11*(9), 1004–1006. https://doi.org/10.1038/nn.2163

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning* (2nd ed.). An Introduction. MIT Press https://play.google.com/store/books/details?id=uWV0DwAAQBAJ.

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews. Neuroscience, 21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

Thome, I., García Alanis, J. C., Volk, J., Vogelbacher, C., Steinsträter, O., & Jansen, A. (2022). Let's face it: The lateralization of the face perception network as measured with fMRI is not clearly right dominant. *Neuroimage, 263*, Article 119587. https://doi.org/10.1016/j.neuroimage.2022.119587

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends Cogn. Sci., 12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Trapp, S., Schweinberger, S. R., Hayward, W. G., & Kovács, G. (2018). Integrating predictive frameworks and cognitive models of face perception. *Psychonomic Bulletin & Review, 25*(6), 2016–2023. https://doi.org/10.3758/s13423-018-1433-x

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). Ffa and OFA encode distinct types of face identity information. *Journal of Neuroscience, 41*(9), 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software, 3*(31), 1026. https://doi.org/10.21105/joss.01026

Vinken, K., Beeck, H. P. O. de, & Vogels, R. (2018). Face repetition probability does not affect repetition suppression in macaque inferotemporal cortex. *Journal of Neuroscience, 38*(34), 7492–7504. https://doi.org/10.1523/JNEUROSCI.0462-18.2018

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vogelbacher, C., Bopp, M. H. A., Schuster, V., Herholz, P., Jansen, A., & Sommer, J. (2019). LAB–QA2GO: A free, easy-to-use toolbox for the quality assessment of magnetic resonance imaging data. *Front. Neurosci., 13*, 688. https://doi.org/10.3389/fnins.2019.00688

Vogelbacher, C., Möbius, T. W. D., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., et al. (2018). The marburg-münster affective disorders cohort study (MACS): A quality assurance protocol for MR neuroimaging data. *Neuroimage, 172*, 450–460. https://doi.org/10.1016/j.neuroimage.2018.01.079

Vogels, R. (2016). Sources of adaptation of inferior temporal cortical responses. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 80*, 185–195. https://doi.org/10.1016/j.cortex.2015.08.024

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

Watkins, C. J. C. H., & Dayan, P. (1992). *Q-learning*. Mach. Learn., *8*(3), 279–292. https://doi.org/10.1007/BF00992698

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *The Behavioral and Brain Sciences, 41*, Article e120. https://doi.org/10.1017/S0140525X17001972