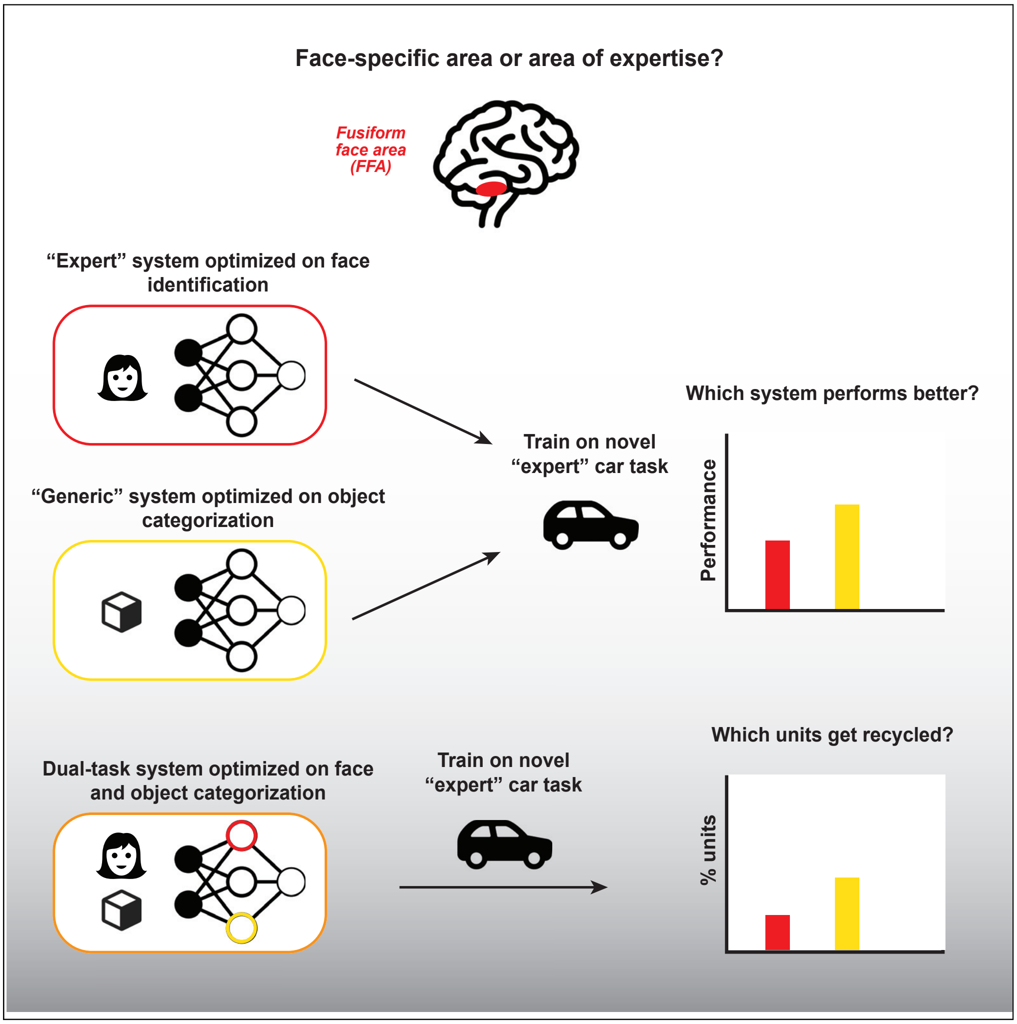
iScience

CellPress
OPEN ACCESS

Article

CNNs reveal the computational implausibility of the expertise hypothesis


Face-specific area or area of expertise?
Fusiform face area (FFA)
"Expert" system optimized on face identification
"Generic" system optimized on object categorization
Train on novel "expert" car task
Which system performs better?
Performance
Dual-task system optimized on face and object categorization
Train on novel "expert" car task
Which units get recycled?
% units

Nancy Kanwisher, Pranjul Gupta, Katharina Dobs

katharina.dobs@psychol.uni-giessen.de

Highlights

Object-trained outperform face-trained CNNs on fine-grained car discrimination

Dual-task CNN uses generic not face-specific features to discriminate cars

A generic "expertise" system spanning categories is computationally implausible

Kanwisher et al., iScience *26*, 105976
February 17, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.105976

iScience

CellPress
OPEN ACCESS

Article

CNNs reveal the computational implausibility of the expertise hypothesis

Nancy Kanwisher,[1,2] Pranjul Gupta,[3] and Katharina Dobs[3,4,1,2,5,*]

SUMMARY

Face perception has long served as a classic example of domain specificity of mind and brain. But an alternative "expertise" hypothesis holds that putatively face-specific mechanisms are actually domain-general, and can be recruited for the perception of other objects of expertise (e.g., cars for car experts). Here, we demonstrate the computational implausibility of this hypothesis: Neural network models optimized for generic object categorization provide a better foundation for expert fine-grained discrimination than do models optimized for face recognition.

INTRODUCTION

Research on face perception has played a central role in cognitive science and neuroscience by providing some of the strongest evidence for the domain specificity of mind and brain. However, some researchers have argued that the cognitive and neural mechanisms engaged during face perception are not specific to faces per se, but play a more general role in the fine-grained discrimination of exemplars of other visual categories for which the person has extensive perceptual expertise (for example, cars for "car experts"[1–4]). Both the behavioral[5] and neural[1] evidence for this alternative "expertise hypothesis"[5] have proven inconsistent at best,[6,7] yet the hypothesis lives on in cognitive neuroscience courses and textbooks.[8] Here, we step back to ask whether the hypothesis that the same neural mechanisms are used for face recognition and discrimination of nonface objects of expertise[1] makes sense computationally in the first place.

A vast body of empirical evidence[9,10] supports the idea that visual recognition is accomplished by the brain through a multi layered hierarchy of perceptual analyzers tuned to increasingly specific visual features, culminating in a neural representation from which linear readout mechanisms can extract the category of an object present in the image.[11] This basic idea is well captured by deep convolutional neural networks (CNNs), which are now capable of near human-like performance on visual recognition tasks.[12] Indeed, considerable evidence shows that these models account for much of the variance in visual recognition behavior and in neural responses recorded from the primate object recognition pathway.[13,14] Although CNNs differ from biological brains in many respects, they offer novel ways of testing optimized solutions to complex real-world computational problems.[15] Here, we use these models to test the computational plausibility of the expertise hypothesis.

RESULTS

An object-trained network outperforms a face-trained network on fine-grained car discrimination

Specifically, we ask whether perceptual expertise with cars would be expected on computational grounds to make use of pre-existing neural machinery for face recognition, or more generic machinery engaged in visual object recognition. We chose cars for this test because they have been widely used in previous work testing the expertise hypothesis,[1,16,17] they generate some of the largest inversion effects seen in nonface stimuli,[18] and they do not have or resemble faces, which has been a concern about some of the prior tests of the expertise hypothesis.[19,20]

We trained the same VGG16 architecture on either face recognition (Face CNN) or object recognition (Object CNN; Figure 1A; cars were excluded from the training set). We then fed images of cars into each network, extracted their representations from the penultimate fully connected layer, and measured how well a linear classifier could decode 100 distinct makes and models of cars based on these image representations. This analysis allows us to quantitatively test the intuition that the features useful for classifying cars

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3] Department of Psychology, Justus-Liebig University Giessen, 35394 Giessen, Germany

[4] Center for Mind, Brain and Behavior (CMBB), University of Marburg and Justus-Liebig University, 35032 Marburg, Germany

[5] Lead contact

*Correspondence: katharina.dobs@psychol.uni-giessen.de

https://doi.org/10.1016/j.isci.2023.105976

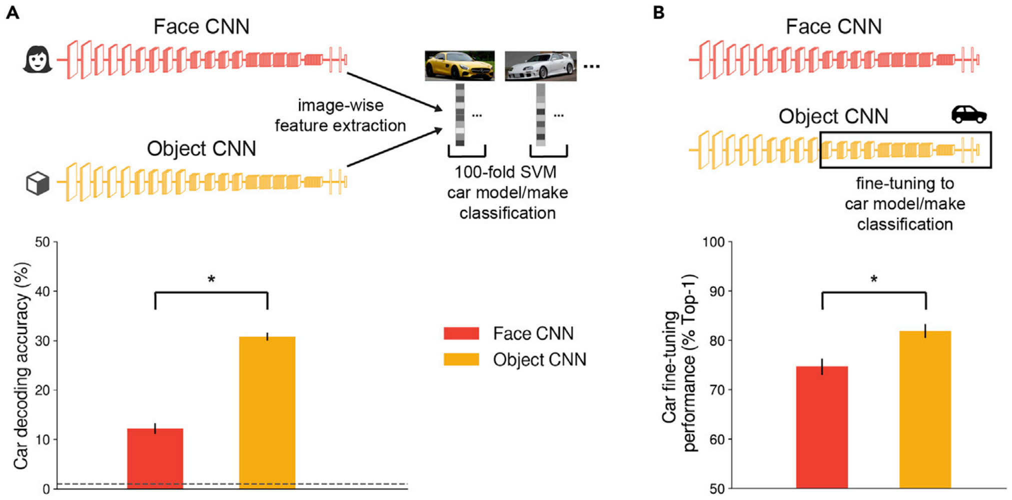


Figure 1. Object-trained CNNs outperform face-trained CNNs on fine-grained car discrimination

(A) We trained deep neural networks with VGG16 architecture on either face identity recognition (Face CNN; red) or object categorization (excluding cars; Object CNN; yellow) and decoded untrained car model/make categories using activation patterns extracted from the penultimate fully connected layer of both CNNs. We found that the Object CNN outperforms the Face CNN in fine-grained car decoding (*p<1e-5, two-sided paired t-test across classification folds). Error bars denote SEM across classification folds. Dashed line indicates chance level (1%). Car images shown are not examples of the original stimulus set due to copyright. Images shown are in public domain and available at https://publicdomainpictures.net.

(B) Performance (% top-1 accuracy on the test set) of the same networks after fine-tuning the CNNs (up to third pooling layer) to a fine-grained car model/make classification task. The Object CNN again outperformed the Face CNN (*p = 0, two-sided bootstrap test). Error bars denote 95% confidence intervals (CIs) bootstrapped across classes and stimuli.

will overlap more with those useful for discriminating objects than with those useful for discriminating faces.[19] Indeed, we found that the object-trained network outperformed the face-trained network (Figure 1A; 31% versus 12%; p<1e-5, two-sided paired t-test across classification folds), suggesting that systems optimized for object recognition work better (right out of the box) than systems optimized for face recognition for fine-grained discrimination of cars.

However, perceptual expertise in humans may develop over many years. Expertise for faces will develop first because newborns are more interested in faces than other objects, and specialized cortical machinery for face perception is already present within the first year of life.[21] Which of the earlier developed systems would then serve as the best foundation on which car expertise could be constructed? To find out, we fine-tuned each network for car discrimination. Specifically, for each network, we removed the original classification layer, replaced it with a new classification layer with 1,109 car model/make classes, and then retrained the network on the car dataset by fine-tuning it from a midlevel layer onwards while freezing the early convolutional layers (up to the third pooling layer). Once performance plateaued we found that car discrimination performance (% top-1 accuracy) was higher on the network originally optimized for object recognition than the network originally optimized for face recognition (Figure 1B; 81% versus 74%; p = 0, two-sided bootstrap test). Again, this finding suggests that expertise with cars can be best accomplished by building on a pre-existing object recognition system, not a pre-existing face recognition system.

A dual-task CNN relies on generic rather than face-specific features to discriminate cars

In actual brains, however, there may be no built-in pipeline that directs training on face recognition (through development or evolution or both) to one system and training on object recognition to another. A more plausible model comes from our recent work in which the same system is trained on both face and object recognition, with no pre-set pathways dividing the two[22] (Figure 2A, left). In that work we found that the dual-trained network spontaneously segregated itself, to increasing degrees at later convolutional layers, into one set of filters that play a greater causal role in face recognition, and another set of filters that play a greater causal role in object recognition. In particular, we showed that some of the filters in

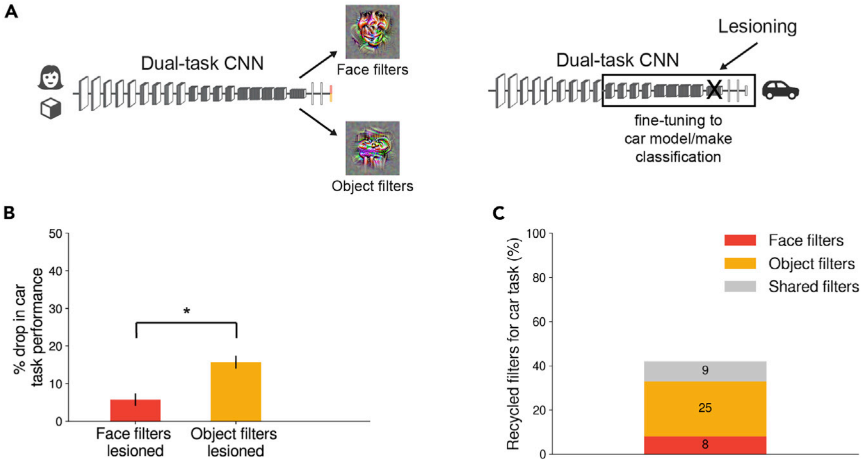


Figure 2. A dual-task CNN relies on generic rather than face-specific features to discriminate cars

(A) The VGG16 network trained on both face and object discrimination (i.e., dual-task CNN, left), where some filters in the last convolutional layer when lesioned produced a greater drop in face recognition than in object categorization (i.e., face filters, top example) performance and others showed the opposite (i.e., object filters, bottom example).[22] We fine-tuned the same network (up to third pooling layer) to a fine-grained car model/make classification task (right) and performed lesioning experiments.

(B) Car discrimination performance of this network drops more when the top 20% of previously identified object filters are lesioned compared to the top 20% of previously identified face filters (*p = 0, two-sided bootstrap test). Error bars denote 95% CIs bootstrapped across classes and stimuli.

(C) The top-20% of filters that maximally harmed the car task when lesioned were composed mostly of filters that were not in the top-20% for either face or object recognition (58% of filters) before fine-tuning for cars, but more overlapped with object (25%) than face (8%) recognition.

the last convolutional layer of the dual-trained network produced a much greater drop in face recognition than object recognition performance when ''lesioned'', and other filters produced the opposite pattern of deficit when lesioned. This enabled us to now ask: When this dual-trained system is subsequently trained on car discrimination, which set of filters takes on that role (the face filters or the object filters)?

To answer this question, we took the network jointly trained on both face and object discrimination and fine-tuned it for car discrimination. We then tested which set of filters produced the greatest drop in car discrimination performance when lesioned (Figure 2A, right). We found that lesioning 20% of filters with the greatest selective role in object recognition reduced accuracy of the car task more than lesioning 20% of filters with the greatest role in face recognition (performance drop of 16% versus 6%; Figure 2B; p = 0, two-sided bootstrap test). Generally, however, the drop in performance was not large in either case. Which filters did the network recycle to perform the car task? To find out, we identified the 20% of filters that maximally harmed the car task when lesioned. Among those filters, only 8% overlapped with face filters, whereas 25% overlapped with object filters and another 9% were recycled from filters that were critical for both face and object tasks (Figure 2C). Of interest, the majority of the filters (58%) were neither face nor object filters, but recycled from the redundant filters in the network. This suggests that car expertise may require its own specialized system, partly recycling a subset of the object system, but largely independent of the face system.

DISCUSSION

Taken together, these findings indicate that systems optimized more broadly for object recognition serve as better foundations than systems optimized for face recognition for subsequent acquisition of car discrimination expertise. Another study conducted independently from ours found similarly that object-trained networks performed better than face-trained networks when retrained for bird discrimination.[23] It would therefore be implausible – though not impossible – for the brain to choose the suboptimal route, using a face-optimized system as the basis for later-learned perceptual expertise.

Here, we have considered one specific version of the ''expertise'' hypothesis, that the very same neural populations are engaged in fine-grained discrimination of faces and of objects of expertise (Note that although fMRI cannot straightforwardly determine whether the same, or merely nearby, neural populations are recruited for different visual categories, results from monkey neurophysiology suggest that the majority of neurons in face-selective patches are indeed face-selective[24]). A less stringent version of the hypothesis merely claims that similar behavioral phenomena may arise for recognition of faces and objects or expertise, an idea we have supported in a different line of work showing that networks trained on car discrimination show face-like inversion effects for cars.[25] Taken together, these findings indicate that perceptual expertise may produce similar processing mechanisms for faces and objects, but those similar mechanisms would not be expected to engage the very same neural populations.

Limitations of the study

Of course we have not tested all possible CNN models, or training regimes, or domains of expertise, and it is possible that our results would differ in other conditions. Furthermore, some of our training choices may not precisely match human experience, such as the number of face categories learned or the ratio of faces to objects for training of the dual-trained network. In addition, impressive as CNNs are in accounting for neural and behavioral data, they are not perfect models of visual object recognition in the brain, as they differ in many respects from biological nervous systems, and it is possible that very different models might make different predictions. For example, a more symbolic or structural description model[26] could in principle account for the perception of both faces and other nonface objects of expertise (but see[27]). However, no such models currently exist that are image-computable and can account for behavioral performance and neural responses at a level even approaching current CNNs. A final caveat is that the human visual recognition system need not be optimal for the recognition of faces and objects because evolution and individual experience optimize for multiple tasks beyond these and descent with modification can produce history-dependent rather than optimal solutions.[28]

More generally, part of the impetus for the expertise hypothesis has been to suggest that if the face system can take on other categories of expertise, then perhaps its ability to process faces is also learned from individual experience, rather than built in innately. On this important broader question of whether the development of the human face system is constrained by innate inductive biases, the current evidence does not yet deliver a clear answer. However, we note that our dual-task network spontaneously segregates itself into distinct face and object systems without any face-specific inductive biases, so at least in principle it is possible for experience alone to create domain-specific systems.[22] In addition, empirical evidence shows that this is clearly true for another nearby cortical region, the visual word form area.[29] On the other hand, empirical challenges that remain to be explained for a purely experiential origin of the face system include its very early development,[21,30] consistent anatomical location, and face-specific activation by touch and sound in congenitally blind people.[31,32] The ultimate answer to this question will inform not just the existence of domain-specific structure in the human brain, but also its developmental and evolutionary origins.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Deep convolutional neural networks
- METHOD DETAILS
  - Training task-specific networks
  - Decoding of fine-grained car discrimination
  - Fine-tuning task-specific CNNs for car discrimination
  - Training and testing a dual-task network
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Significance testing

ACKNOWLEDGMENTS

We thank Julio Martinez for his help and advice in programming the experiments, members of the Kanwisher laboratory for fruitful discussions and feedback and Rebecca Saxe for valuable comments on the manuscript. This work was supported by DFG Collaborative Research Center SFB/TRR 135 and by the Research Cluster "The Adaptive Mind", funded by the Hessian Ministry for Higher Education, Research Science and the Arts to K.D., NIH grant DP1HD091947to N.K and National Science Foundation Science and Technology Center for Brains, Minds, and Machines.

AUTHOR CONTRIBUTIONS

N.K. and K.D. conceived the study. P.G. and K.D. performed the analyses. N.K., K.D., and P.G. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 13, 2022
Revised: November 7, 2022
Accepted: January 11, 2023
Published: February 17, 2023

REFERENCES

1. Gauthier, I., Skudlarski, P., Gore, J.C., and Anderson, A.W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. Nat. Neurosci. *3*, 191–197. https://doi.org/10.1038/72140.

2. Gauthier, I., Tarr, M.J., Anderson, A.W., Skudlarski, P., and Gore, J.C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. Nat. Neurosci. *2*, 568–573. https://doi.org/10.1038/9224.

3. Tarr, M.J., and Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nat. Neurosci. *3*, 764–769. https://doi.org/10.1038/77666.

4. Ross, D.A., Tamber-Rosenau, B.J., Palmeri, T.J., Zhang, J., Xu, Y., and Gauthier, I. (2018). High-resolution functional magnetic resonance imaging reveals configural processing of cars in right anterior fusiform face area of car experts. J. Cogn. Neurosci. *30*, 973–984. https://doi.org/10.1162/jocn_a_01256.

5. Diamond, R., and Carey, S. (1986). Why faces are and are not special: an effect of expertise. J. Exp. Psychol. Gen. *115*, 107–117. https://doi.org/10.1037/0096-3445.115.2.107.

6. Grill-Spector, K., Knouf, N., and Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. Nat. Neurosci. *7*, 555–562. https://doi.org/10.1038/nn1224.

7. Op de Beeck, H.P., Baker, C.I., DiCarlo, J.J., and Kanwisher, N.G. (2006). Discrimination training alters object representations in human extrastriate cortex. J. Neurosci. *26*, 13025–13036. https://doi.org/10.1523/jneurosci.2481-06.2006.

8. Postle, B.R. (2020). Essentials of Cognitive Neuroscience, 2nd edition (Wiley & Sons).

9. Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. J. Physiol. *195*, 215–243.

10. Van Essen, D.C., and Maunsell, J.H. (1983). Hierarchical organization and functional streams in the visual cortex. Trends Neurosci. *6*, 370–375. https://doi.org/10.1016/0166-2236(83)90167-4.

11. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866. https://doi.org/10.1126/science.1117593.

12. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. NIPS. 1097–1105.

13. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.

14. Ratan Murty, N.A., Bashivan, P., Abate, A., DiCarlo, J.J., and Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat. Commun. *12*, 5540. https://doi.org/10.1038/s41467-021-25409-6.

15. Kanwisher, N., Khosla, M., and Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. Trends Neurosci *46*, 1883. https://doi.org/10.1016/j.tins.2022.12.008.

16. Xu, Y., Liu, J., and Kanwisher, N. (2005). The M170 is selective for faces, not for expertise. Neuropsychologia *43*, 588–597. https://doi.org/10.1016/j.neuropsychologia.2004.07.016.

17. Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. Cereb. Cortex *15*, 1234–1242. https://doi.org/10.1093/cercor/bhi006.

18. Rezlescu, C., Chapman, A., Susilo, T., and Caramazza, A. (2016). Large inversion effects are not specific to faces and do not vary with object expertise. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/xzbe5.

19. Kanwisher, N. (2000). Domain specificity in face perception. Nat. Neurosci. *3*, 759–763. https://doi.org/10.1038/77664.

20. Lochy, A., Zimmermann, F.G.S., Laguesse, R., Willenbockel, V., Rossion, B., and Vuong, Q.C. (2018). Does extensive training at individuating novel objects in adulthood lead to visual expertise? The role of facelikeness. J. Cogn. Neurosci. *30*, 449–467. https://doi.org/10.1162/jocn_a_01212.

21. Kosakowski, H.L., Cohen, M.A., Takahashi, A., Keil, B., Kanwisher, N., and Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. Curr. Biol. *32*, 265–274.e5. https://doi.org/10.1016/j.cub.2021.10.064.

22. Dobs, K., Martinez, J., Kell, A.J.E., and Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Sci. Adv. *8*, eabl8913. https://doi.org/10.1126/sciadv.abl8913.

23. Yovel, G., Grosbard, I., and Abudarham, N. (2022). Deep learning models of perceptual expertise support a domain-specific account. Preprint at bioRxiv. https://doi.org/10.1101/2022.12.01.518342.

24. Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674. https://doi.org/10.1126/science.1119983.

25. Dobs, K., Yuan, J., Martinez, J., and Kanwisher, N. (2022). Using deep convolutional neural networks to test why human face recognition works the way it does. Preprint at bioRxiv. https://doi.org/10.1101/2022.11.23.517478.

26. Marr, D. (1982). Vision (Freeman).

27. Biederman, I., and Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. Philos. Trans. R. Soc. Lond. B Biol. Sci. *352*, 1203–1219. https://doi.org/10.1098/rstb.1997.0103.

28. Marcus, G. (2009). Kluge: The Haphazard Evolution of the Human Mind (Houghton Mifflin Harcourt).

29. McCandliss, B.D., Cohen, L., and Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. Trends Cogn. Sci. *7*, 293–299. https://doi.org/10.1016/s1364-6613(03)00134-7.

30. Farroni, T., Johnson, M.H., Menon, E., Zulian, L., Faraguna, D., and Csibra, G. (2005). Newborns' preference for face-relevant stimuli: effects of contrast polarity. Proc. Natl. Acad. Sci. USA *102*, 17245–17250. https://doi.org/10.1073/pnas.0502205102.

31. Ratan Murty, N.A., Teng, S., Beeler, D., Mynick, A., Oliva, A., and Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. Proc. Natl. Acad. Sci. USA *117*, 23011–23020. https://doi.org/10.1073/pnas.2004607117.

32. van den Hurk, J., Van Baelen, M., Op de Beeck, H.P., and Hans. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. Proc. Natl. Acad. Sci. USA *114*, E4501–E4510. https://doi.org/10.1073/pnas.1612862114.

33. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. https://doi.org/10.1109/cvpr.2009.5206848.

34. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., and Zisserman, A. (2018). VGGFace2: a dataset for recognising faces across pose and age. In IEEE International Conference on Automatic Face & Gesture Recognition IEEE International Conference on Automatic Face & Gesture Recognition, pp. 67–74.

35. Yang, L., Luo, P., Loy, C.C., and Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3973–3981. https://doi.org/10.1109/cvpr.2015.7299023.

36. Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations International Conference on Learning Representations, pp. 1–14.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ImageNet dataset	ImageNet: A large-scale hierarchical image database[33]	https://www.image-net.org/challenges/LSVRC/
VGGFace2 dataset	VGGFace2: A dataset for recognising faces across pose and age[34]	https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/
CompCars dataset	A Large-Scale Car Dataset for Fine-Grained Categorization and Verification[35]	http://mmlab.ie.cuhk.edu.hk/datasets/comp_cars/index.html
Software and algorithms		
Untrained VGG16	VGG16 Model[36]	https://www.pytorch.org
Pytorch version 3.x	Meta AI	https://pytorch.org/
Numpy 1.x	Community project	https://numpy.org/
OpenCV 4.x	Intel Corporation, Willow Garage, Itseez	https://opencv.org/
OSF	Center for Open Science	https://osf.io/
Scipy 1.x	Travis Oliphant, Pearu Peterson, Eric Jones	https://scipy.org/
sdnn	Julio Martinez	https://github.com/martinezjulio/sdnn
Matplotlib 3.x	John D. Hunter	https://matplotlib.org/
Other		
NVIDIA GPUs	Nvidia Corporation	https://www.nvidia.com/en-us/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for code and resources should be directed to and will be fulfilled by the lead contact, Katharina Dobs (katharina.dobs@psychol.uni-giessen.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The paper used existing datasets and code to train the computational models. The original datasets and codes are listed in the key resources table.
- Data and code to generate the plots are available at the Open Science Framework repository for this project (https://osf.io/qtnzs/). Additional code to fine-tune custom models can be found at the Github repository: https://github.com/kathadobs/exphypo.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Deep convolutional neural networks

To test whether a generic or a face-specific system would serve as better foundation for learning another fine-grained task, we trained two randomly initialized deep convolutional neural networks with a VGG16 architecture[36] on object categorization (Object CNN) and on face identity categorization (Face CNN).

To further test which of the two systems – if both are available simultaneously – would be recycled for learning a new fine-grained discrimination task, we used a previously developed dual-task network which had been trained on faces and objects. The details for training this network have been explained previously.[22]

METHOD DETAILS

Training task-specific networks

To train the Object CNN, we manually sampled 423 prototypical object categories (e.g., trumpet, hammer, coffee cup) of the ILSVRC-2012 database[33] and used these images for training. In particular, to avoid including another fine-grained distinction within the object dataset, we removed all categories from the original dataset that were animals (e.g., ImageNet contains many bird and dog species). Critically, we further removed all vehicle-related categories. For each of the 423 selected object categories, we used 1000 images for training and 200 for validation, for a total of 423,000 training and 84,600 validation images. For training, we used stochastic gradient descent (SGD) with momentum with an initial learning rate of 10^{-3}, a weight decay of 10^{-4} and momentum of 0.9. We manually reduced the learning rate twice to 10^{-4} and 10^{-5} when the training loss did not decrease for five epochs (i.e., full passes over the training set). To update the weights during training, we computed the cross-entropy loss on random batches of 128 object images and backpropagated the loss. Each image was scaled to a minimum side length (height or width) of 256 pixels, normalized to a mean and SD of 0.5, and data augmentation (i.e., 20% gray-scaled, randomly cropped to 224 × 224 pixel) was applied during training. The test images were scaled, normalized and center-cropped before extracting the classification.

To train the Face CNN, we used 1,714 identities (857 female) from the VGGFace2 database.[34] We chose identities with a minimum of 300 images per identity and balanced female and male identities (857 each), otherwise the identities were randomly selected from the VGGFace2 database. To match the training set size to the Object CNN, we randomly chose 246 images per face identity for training, and 50 images per category for validation, for a total of 421,644 training and 85,700 validation images. All other learning parameters were identical to the Object CNN.

Decoding of fine-grained car discrimination

To test whether the representations optimized for face or generic object categorization would be more useful for fine-grained discrimination of exemplars of another stimulus category, we decoded exemplars of car model/make categories from the activations extracted from both networks. We used 100 car model/make categories (10 each; 1000 images total) from the CompCars dataset.[35] We extracted the activation in the penultimate fully-connected layer of each network (i.e., the last layer before the classification layer) to the 1000 car images. For activations from each network, we trained and tested a 100-way linear support vector machine (with L2 regularization) on the corresponding activation patterns using a leave-one-image-out (i.e., 10-fold) cross-validation scheme.

Fine-tuning task-specific CNNs for car discrimination

In human experience, becoming an expert requires learning over time. To test whether the object- or the face-trained network will serve as a better foundation to learn another fine-grained discrimination task, we fine-tuned each network to a car model/make classification task. We used 1,109 classes with 45 images for training and 5 images for testing per class, for a total of 49,905 training and 5,545 validation images from the CompCars dataset. To obtain enough images per class, we concatenated images from the same model/make but of different years into one class. For each network, we removed the original classification layer (e.g., 1,714 face classes for the Face CNN) and replaced it with a new classification layer with 1,109 car model/make classes. We then retrained each network on the car dataset by fine-tuning it from a midlevel layer onwards, while freezing the early convolutional layers (up to the third pooling layer). We decided to freeze the early convolutional layers since we did not expect expertise to change early visual processing. Indeed, we found no significant difference in car decoding accuracy between the face-trained and the object-trained CNN based on representations extracted from the third pooling layer (24% vs. 23%; $p > 0.8$, two-sided paired t-test across classification folds). Except for those early layers, all other filters were free to change during training allowing for plasticity throughout the network. For fine-tuning, we used SGD with momentum with an initial learning rate of 10^{-3}, a weight decay of 10^{-4} and momentum of 0.9. To update the weights of the fully-connected layers, we computed the cross-entropy loss on random batches of

128 object images and backpropagated the loss. The top-1% accuracy on the test set served as performance for each of the networks. We obtained bootstrapped 95% confidence intervals (CIs) for the accuracy of both networks by bootstrapping and computing the accuracy across classes and images 10,000 times.

Training and testing a dual-task network

Humans typically do not become experts of another fine-grained task until at least several years of age, at which point development of their face and object recognition system is already quite advanced. If both systems are available simultaneously, which of these two systems would be recycled for learning a new fine-grained discrimination task? To test this, we used a dual-task network trained on face recognition and object categorization.[22] Importantly, this network spontaneously developed two sets of filters in the last convolutional layer that were predominantly involved in either face or object recognition, while few filters were shared between both tasks. Here, we wanted to find out which of these two subsystems would be recycled when learning a new fine-grained discrimination task. To test this, we first retrained the dual-task network to the same car model/make discrimination task as used for the face-trained and the object-trained CNN. We again fine-tuned layers of the dual-task CNN from a midlevel layer onwards (while freezing the early convolutional layers up to the third pooling layer), using the identical dataset and parameters for training. The top-1% accuracy on the test set served as base performance for the car task in this network. To find out which of the two original subsystems (face or object filters) were taken up by the car task, we then lesioned the entire set of the previously identified 20% highest-ranked filters (i.e., face or object filters) in the last convolutional layer of the CNN and computed the % drop in performance on the car task based on the validation set. We obtained bootstrapped 95% confidence intervals (CIs) for the % drop in accuracy for both tasks by bootstrapping and computing the drop in accuracy across classes and images 10,000 times.

To identify which filters were causally involved in the car task, we performed lesioning experiments in the fine-tuned dual-task network. First, we identified putative car-specific filters in the last convolutional layer by evaluating how much ablating that filter (i.e., setting its output to zero) affected the loss for 50 batches of car images, all taken from the training set. We then ranked the filters by how much they affected the loss on car images. Using a greedy procedure, we first selected and dropped the highest-ranking group (~1.6%) of filters, then selected the next highest-ranking group from the remaining filters in similar fashion but on novel batches of images (also taken from the training set). We repeated this process until there were no remaining filters left, resulting in all filters being ranked for the impairment they produced on the car task when lesioned. For the 20% highest ranked filters (i.e., 102 filters), we then computed the % overlap of these filters with the 20% highest face- and the 20% highest object-ranked filters. In particular, we computed the overlap for filters that belonged to the highest-ranked filter of the face task but not the object task (face filters), filters that ranked high on the object but not the face task (object filters), and filters that ranked high on both the face and the object task (shared filters).

QUANTIFICATION AND STATISTICAL ANALYSIS

Significance testing

For SVM decoding accuracies, we computed the mean and SEM across classification folds and used two-sided paired t-tests across classification folds to test for differences between decoding accuracies. Significance of comparisons between the top-1% accuracies of different networks was obtained by using direct bootstrap tests. In particular, for statistical inference of the differences between performances (or drop in performances), we bootstrapped the classes and images 10,000 times and computed the mean difference between accuracies (or % drop in accuracies) resulting in an empirical distribution of performance differences. The minimum number of differences that were smaller or larger than zero divided by the number of bootstraps defined the p value (i.e., two-sided testing). Primarily, NumPy has been used for data analysis. The statistical parameters (p values along with required tests) are reported alongside their descriptions in the main text of the paper.