Dr. Robert Geirhos

✉ Email	lastname@google.com	Address	Google DeepMind
Website	https://robertgeirhos.com		Brandschenkestrasse 110
GitHub	https://github.com/rgeirhos		8002 Zürich
LinkedIn	https://linkedin.com/in/rgeirhos		Switzerland
Scholar	Robert Geirhos		

Professional experience

Google DeepMind

Since 5/26 Staff Research Scientist, *Google DeepMind*

5/25 – 4/26 Senior Research Scientist, *Google DeepMind*

8/22 – 4/25 Research Scientist, *Google Brain, later Google DeepMind*

University of Tübingen — 🎓 Felix Wichmann

11/21 – 6/22 Postdoctoral Researcher, *Department of Computer Science* — 🎓 Wieland Brendel

Meta AI & Facebook AI Research (FAIR)

1/22 – 5/22 External Research Collaborator, *part-time via PRO Unlimited* — 🎓 Ari Morcos

7/21 – 10/21 PhD Research Intern — 🎓 Surya Ganguli

University of Tübingen & IMPRS-IS

11/18 – 7/21 Doctoral Researcher, *Department of Computer Science & International Max Planck Research School for Intelligent Systems* — 🎓 Felix Wichmann, 🎓 Matthias Bethge

Education

2018 – 2022 **Ph.D. in Computer Science, summa cum laude**
German title: Dr. rer. nat. (Doctor of Science)
Department of Computer Science, University of Tübingen & International Max Planck Research School for Intelligent Systems

2016 – 2018 **M.Sc. Computer Science, with distinction**
Grade 1.1/1.0 (“excellent”)
Department of Computer Science, University of Tübingen

2017 – 2018 Exchange student
School of Informatics, University of Amsterdam

2013 – 2016 **B.Sc. Cognitive Science**
Grade 1.37/1.0 (“excellent”)
Department of Computer Science, University of Tübingen

2015 Exchange student
Schools of Computing Science & Psychology, University of Glasgow

2004 – 2012 **Abitur**
Grade 1.0/1.0 (“excellent”)
Montfort-Gymnasium Tettnang

"Weltwärts" development volunteer service

2012 – 2013 Live-in volunteer (12 months)
Asha Niketan is a L'Arche interfaith community in Kolkata, India, where people with and without intellectual disabilities not only work together, but also share their life.
Asha Niketan Kolkata (formerly Calcutta), India

Open Source projects

Contributing to Open Source is important to me. I created and maintain the following Open Source projects, either as the lead developer (★) or as the lead supervisor (☆):

★ 800+ rgeirhos/texture-vs-shape: an analysis method to assess the degree to which black-box neural network models rely on local (texture) features.

★ 500+ rgeirhos/Stylized-ImageNet: code to create Stylized-ImageNet, a strong data augmentation technique for increasing out-of-distribution robustness.

☆ 450+ bethgelab/imagecorruptions: a Python package for out-of-distribution robustness testing, used by 4,000+ repos as part of mmdetection, a leading object detection toolbox.

★ 350+ bethgelab/model-vs-human: a Python toolbox to test the out-of-distribution robustness and biases of PyTorch and TensorFlow models against high-quality human data.

☆ 250+ google-deepmind/physics-IQ-benchmark: an evaluation protocol, benchmark and dataset for evaluating physical understanding in generative video models.

★ 150+ bethgelab/robust-detection-benchmark: a benchmark to assess the robustness of object detection models towards outdoor hazards like snow / fog / frost.

☆ 150+ bethgelab/stylize-datasets: a script to stylize arbitrary datasets through Adaptive Instance Normalization (AdaIN).

☆ 100+ rgeirhos/shortcut-perspective: Python code to reproduce shortcut learning in neural networks.

★ 100+ rgeirhos/academic-cv-publications: a simple LaTeX template to generate a highly customised list of publications using BibTeX entries (developed for this very CV).

★ 50+ rgeirhos/generalisation-humans-DNNs: code and psychophysical data for 16-class-ImageNet, a dataset for comparing human and machine classification decisions.

★ 50+ rgeirhos/dataset-pruning-metris: metrics for smart dataset pruning, resulting in training efficiency gains (and thus carbon savings) of about 20% without loss of performance.

★ 25+ google-research/fooling-feature-visualizations: adversarial model manipulation technique to fool feature visualizations, a widely used interpretability tool.

★ 10+ google-deepmind/visual-memory: fast nearest neighbor search and analysis techniques including memory pruning to improve a deep neural network based visual memory.

Talks

06/2026 Neural Information Processing Group, University of Tübingen (upcoming)

06/2026 Keynote at CVPR workshop on Visual General Intelligence, Denver (upcoming)

03/2026 Lassonde School of Engineering, York University

03/2026 ELLIS | ELIAS | ELLIOT Winterschool on Foundation Models, Amsterdam
12/2025 Enigma Journal Club, Stanford
11/2025 KUIS AI Talks, Koç University
08/2025 Google DeepMind horizon talk, Mountain View
05/2025 Keynote at Boehringer Ingelheim Data Science Conference, Berlin
04/2025 Kriegeskorte Lab, Columbia University
12/2024 Keynote at NeurIPS Workshop on Attributing Model Behavior at Scale, Vancouver
11/2024 Keynote at BMVC Workshop on Robust Recognition in the Open World, Glasgow
10/2024 Centre for Integrative and Applied Neuroscience, York University
09/2024 Waymo, Mountain View
05/2024 VSS 2024 Oral, St. Pete Beach
02/2024 Google DeepMind research week plenary session, Mountain View
01/2024 Cognitive Science Speaker Series, York University
11/2023 Google DeepMind foundational research summit, London
11/2023 Gatsby Computational Neuroscience Unit, University College London
10/2023 Interpretability Analysis Community Meeting, London
09/2023 Vision Seminar, University of Illinois Urbana-Champaign
07/2023 Clinical AI Journal Club, University College London
07/2023 German Dissertation Award Finalists, Körber-Stiftung
05/2023 Gesellschaft für Informatik Dissertation Award Finalists, Schloss Dagstuhl
06/2022 RIKEN-AIP TrustML seminar, Tokyo
06/2022 Institute for Adaptive and Neural Computation, University of Edinburgh
05/2022 NYU Vision Journal Club, New York University & Flatiron Institute
05/2022 VSS 2022 Oral, St. Pete Beach
04/2022 NCT Data Science Seminar, DKFZ (recording)
03/2022 Neural Networks Colloquium, University of Tübingen
02/2022 Apple Machine Learning Research Seminar, Cupertino
12/2021 NeurIPS 2021 Oral, Sydney
09/2021 Brain & AI meeting, Facebook AI Research, New York
12/2020 NeurIPS 2020 SVRHM Workshop Oral (recording)
07/2020 AI Lunch Series, Jefferson National Laboratory
07/2020 Current Topics in Perception and Cognition Colloquium, Giessen University
05/2019 VSS 2019 Oral, St. Pete Beach
05/2019 Center for Perceptual Systems, University of Texas at Austin
05/2019 ICLR 2019 Oral, New Orleans (recording)
03/2019 AI Meetup, Hamburg
08/2017 Neural Networks Colloquium, University of Tübingen
08/2016 Perceiving Systems, Max Planck Institute for Intelligent Systems

Awards & Scholarships

2023	ELLIS PhD Award for outstanding European machine learning dissertation | *ELLIS*
2022	Outstanding Paper Award for "Beyond neural scaling laws: beating power law scaling via data pruning" | *NeurIPS*
2022	Dissertation Award | *University of Tübingen*
2013 – 2022	e-fellows.net Scholarship | *Awarded for outstanding academic achievements & engagement*
2018 – 2021	Ph.D. Scholarship | *International Max Planck Research School for Intelligent Systems*
2020	Elsevier/Vision Research Travel Award
2019	ICLR Travel Award
2018	NeurIPS Travel Award
2017	German National Scholarship (Deutschlandstipendium) | *University of Tübingen*
2017	Erasmus+ Mobility Stipend | *University of Amsterdam*
2015	Erasmus+ Mobility Stipend | *University of Glasgow*
2015	Lilli Zapf Award | *Acknowledges commitment to social initiative and tolerance*
2012	Award of the City of Tettnang | *Acknowledges outstanding academic achievements*

Publications

Peer-reviewed articles

CAIS, ScaleAI, & HLE-Contributors-Consortium. (2026). A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, *649*(8099), 1139–1146.

Motamed, S., Culp, L., Swersky, K., Jaini, P., & Geirhos, R. (2026). Do generative video models understand physical principles? In *Proceedings of the Winter Conference on Applications of Computer Vision*. [code | website]

Hewitt, J., Tafjord, O., Geirhos, R., & Kim, B. (2026). Neologism Learning for Controllability and Self-Verbalization. In *International Conference on Learning Representations*.

[Spotlight] Hewitt, J., Geirhos, R., & Kim, B. (2025). We Can't Understand AI Using our Existing Vocabulary. In *International Conference on Machine Learning*.

Internò, C., Geirhos, R., Olhofer, M., Liu, S., Hammer, B., & Klindt, D. (2025). AI-generated video detection via perceptual straightening. In *Advances in Neural Information Processing Systems 38*. [code]

Geirhos, R., Jaini, P., Stone, A., Medapati, S., Yi, X., Toderici, G., … Shlens, J. (2025). Towards flexible perception with visual memory. In *International Conference on Machine Learning*. [code]

Li, F., Klein, T., Brendel, W., Geirhos, R., & Zimmermann, R. S. (2025). LAION-C: An out-of-distribution benchmark for web-scale vision models. In *International Conference on Machine Learning*. [code]

Gavrikov, P., Lukasik, J., Jung, S., Geirhos, R., Lamm, B., Mirza, M. J., … Keuper, J. (2025). Can we talk models into seeing the world differently? In *International Conference on Learning Representations*. [code]

Stone, A., Soltau, H., Geirhos, R., Yi, X., Xia, Y., Cao, B., … Shlens, J. (2025). Learning Visual Composition through Improved Semantic Guidance. In *Conference on Computer Vision and Pattern Recognition.*

[Spotlight] Jaini, P., Clark, K., & Geirhos, R. (2024). Intriguing properties of generative classifiers. In *International Conference on Learning Representations.*

Ahlert, J., Klein, T., Wichmann, F. A., & Geirhos, R. (2024). How aligned are different alignment metrics? In *ICLR 2024 Workshop on Representational Alignment.*

Geirhos, R., Zimmermann, R. S., Bilodeau, B., Brendel, W., & Kim, B. (2024). Don't trust your eyes: on the (un) reliability of feature visualizations. In *International Conference on Machine Learning.* [code]

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., … others (2023). Getting aligned on representational alignment. *Transactions on Machine Learning Research.*

Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., … others (2023). Patch n'Pack: NaViT, a vision transformer for any aspect ratio and resolution. In *Advances in Neural Information Processing Systems 36.*

Wichmann, F. A., Kornblith, S., & Geirhos, R. (2023). Neither hype nor gloom do DNNs justice. *Behavioral and Brain Sciences*, *46*, e412.

[Oral] Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., … others (2023). Scaling vision transformers to 22 billion parameters. In *International Confererence on Machine Learning.*

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, *9*.

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2023). The developmental trajectory of object recognition robustness: Children are like small adults but unlike big deep neural networks. *Journal of Vision*, *23*(7). [code]

[Award] Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., & Morcos, A. S. (2022). Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems 35.* [code | data]

Meding, K., Buschoff, L. M. S., Geirhos, R., & Wichmann, F. A. (2022). Trivial or impossible—dichotomous data difficulty masks model differences (on ImageNet and beyond). *International Conference on Learning Representations.* [code]

[Oral] Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems 34.* [code]

[Spotlight] Zimmermann, R. S., Borowski, J., Geirhos, R., Bethge, M., Wallis, T. S. A., & Brendel, W. (2021). How well do feature visualizations support causal understanding of CNN activations? In *Advances in Neural Information Processing Systems 34.* [code]

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2021). Out-of-distribution robustness: Limited image exposure of a four-year-old is enough to outperform ResNet-50. In *NeurIPS Workshop on Shared Visual Representations in Human & Machine Intelligence.* [code]

Borowski, J., Zimmermann, R., Schepers, J., Geirhos, R., Wallis, T. S. A., Bethge, M., & Brendel, W. (2021). Exemplary natural images explain CNN activations better than feature visualizations. *International Conference on Learning Representations.*

[**Oral**] Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F. A., & Brendel, W. (2020). On the surprising similarities between supervised and self-supervised models. In *NeurIPS Workshop on Shared Visual Representations in Human & Machine Intelligence.*

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020b). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673. [code]

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Advances in Neural Information Processing Systems 33.* [code]

[**Oral**] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations.* [models | dataset]

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., … Brendel, W. (2019). Benchmarking robustness in object detection: autonomous driving when winter is coming. In *NeurIPS Workshop on Machine Learning for Autonomous Driving.* [benchmark | library]

Geirhos, R., Medina Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems 31* (pp. 7548–7560). [code]

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, Human Vision and Electronic Imaging*, *2017*(14), 36–45.

Preprints

Wiedemer, T., Li, Y., Vicol, P., Gu, S. S., Matarese, N., Swersky, K., … Geirhos, R. (2025). Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*. [website]

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., … others (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261.*

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., … others (2025). Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*. [website | Wikipedia article]

Haghiri, S., Rubisch, P., Geirhos, R., Wichmann, F., & von Luxburg, U. (2019). Comparison-based framework for psychophysics: lab versus crowdsourcing. *arXiv preprint arXiv:1905.07234.*

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*. [code]

Conference abstracts

[**Oral**] Geirhos, R., Clark, K., & Jaini, P. (2024). Learning to discriminate by learning to generate: zero-shot generative models increase human object recognition alignment. *Journal of Vision*, *24*(10).

[**Oral**] Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2022). The bittersweet lesson: data-rich models narrow the behavioural gap to human vision. *Journal of Vision*, *22*(14), 3273–3273.

[Oral] Huber, L., Geirhos, R., & Wichmann, F. A. (2021). The developmental trajectory of object recognition robustness: comparing children, adults, and CNNs. *Journal of Vision*, *21*(9), 1967.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020a). Unintended cue learning: Lessons for deep learning from experimental psychology. *Journal of Vision*, *20*(11), 652–652.

[Oral] Geirhos, R., Rubisch, P., Rauber, J., Medina Temme, C. R., Michaelis, C., Brendel, W., … Wichmann, F. A. (2019). Inducing a human-like shape bias leads to emergent human-level distortion robustness in CNNs. *Journal of Vision*, *19*(10), 209c–209c.

Geirhos, R., Janssen, D., Schütt, H., Bethge, M., & Wichmann, F. (2017). Of human observers and deep neural networks: A detailed psychophysical comparison. *Journal of Vision*, *17*(10), 806–806.

Butz, M. V., Geirhos, R., & Kneissler, J. (2015). An automatized Heider-Simmel story generation tool. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci), Pasadena, California, USA, July 22-25, 2015.* [website & code]

Doctoral dissertation

[Award] Geirhos, R. (2022). To err is human? A functional comparison of human and machine decision-making. *University of Tübingen.*

Recognized with the European ELLIS PhD award, the University of Tübingen's 2022 dissertation award, and named an "outstanding dissertation" by the German Society for Computer Science (GI).

Selected media coverage

My work was featured in print (The Economist, New York Times, FAZ, Spektrum der Wissenschaft), online (e.g. Quanta Magazine, Knowable Magazine), radio (SWR), YouTube (DeepMind's AI Just Solved Video Generation In A Way Nobody Expected [257K views], Do Neural Networks Need To Think Like Humans? [48K views], Finally, DeepMind made an IQ test for AIs! [65K views]), & podcast (Underrated ML).

Additionally, I have written articles for The Gradient (Shortcuts: How Neural Networks Love to Cheat), Towards Data Science (Are all CNNs created equal?), Medium (Out of shape? Why deep learning works differently than we thought) and the Machine Learning for Science Blog (Do machines see like humans? They are getting closer).

Professional activities & Service

Area Chair	NeurIPS, ICML, ICLR
Reviewer	NeurIPS (*Outstanding Reviewer Award*, 2020), ICLR, ICML, Nature Machine Intelligence, Nature Human Behaviour, Journal of Vision (2×*Exceptionally Good Review*), PLOS Computational Biology, ICCV, WACV, ICLR 2024–2025 workshop on Representational Alignment, CVPR 2024 workshop on Test-Time Adaptation, NeurIPS 2022 workshop on Distribution Shifts, ICML 2020 workshop on Uncertainty & Robustness in Deep Learning, NeurIPS 2019–2022 workshop on Shared Visual Representations in Humans & Machines, Computer Vision and Image Understanding, AISTATS (emergency reviewer)
Organizer	ICCV Physics-IQ × Perception Test Challenge, 2025
Assessor	External examiner, dissertation of Thomas Fel | Toulouse/Brown Grant reviewer & subject-matter expert for MacArthur Foundation ($800K / fellowship) Grant reviewer for UK Research & Innovation (£2M / grant) Grant reviewer for Israeli Science Foundation (ILS 1M / grant) Grant reviewer for CAHSI-Google Institutional Research Program ($80K / grant) Annual project evaluator for BWKI, Germany's national AI school competition
Member	ELLIS Society
Committees	Professorship Appointment Committee "Computer Science & Didactics", Tübingen Professorship Interim Evaluation Committee "Big Data Visual Analytics", Tübingen
Panelist	Attributing Model Behavior at Scale, NeurIPS 2024, Vancouver Career transitions panel, VSS 2024, St. Pete Beach, Florida "Artificial stupidity? On accidents and deceptions of technical intelligence", Dresden
Teaching	Annual tutorial on Rebuttal Writing | Max Planck Research School for Intelligent Systems Interdisciplinary seminar on Artificial Intelligence & Legal Tech | University of Tübingen
Mentor	FoVea network (Females of Vision et al.), 2024

Mentoring & Supervision

Interns & Lab rotat.	Thaddäus Wiedemer | Google DeepMind Toronto *"Video models are zero-shot learners and reasoners"*
	Saman Motamed | Google DeepMind Toronto *"Do generative video models understand physical principles?"*
	Ole Jonas Wenzel | GTC Tübingen *"Imperceptible signals in perceptible noise"*
	Shuchen Wu | ETH Zürich *"An early vision-inspired visual recognition model"*
	Benjamin Mitzkus | University of Tübingen *"Benchmarking robustness in object detection"*
	Patricia Rubisch | University of Tübingen *"Comparison-based framework for psychophysics: lab versus crowdsourcing"*
M.Sc. theses	Fanfei Li | Max Planck Institute for Intelligent Systems *"LAION-C: An out-of-distribution benchmark for web-scale vision models"*
	Lukas Huber | University of Bern *"An onto- and phylogenetically inspired approach to machine vision"*
	Benjamin Mitzkus | University of Tübingen *"Meta-learning robustness"*
B.Sc. theses	Jannis Ahlert | University of Tübingen *"How aligned are different alignment metrics?"*
	Tizian Thieringer | University of Tübingen *"Benchmarking the latest machine vision developments against human categorization performance"*

Languages

English	full professional proficiency | I've worked in English-speaking environments for a decade.
German	native speaker | Deutsch ist meine Muttersprache.
French	good command | Je parlais très bien le français, mais maintenant j'ai peu de pratique.
Bengali	basic | আমি একবছর কলকাতায় বাংলা বলতে এবং লিখতে শিখেছি।
Spanish	basic | Estoy aprendiendo algo de español.