

Martina G. Vilas

🌐 martinagvilas.github.io

@ martinagonzalezvilas@gmail.com

🔄 github.com/martinagvilas

🎓 scholar.google.com/user=X_n9N

ABOUT ME

I am a computer science doctoral researcher with a background in cognitive neuroscience. Working at the intersection of these topics, my research focuses on reverse engineer the cognitive capacities of AI models and improve their alignment with human cognition.

EDUCATION

Doctoral degree in Computer Science | Goethe University

Thesis topic in the field of inner interpretability of AI models. Co-supervised by Prof. Gemma Roig and Prof. David Poeppel.

- Passed qualifying exam in *Theoretical Computer Science, Software Engineering and Hardware*.

ongoing
Germany

Licenciatura in Psychology, with a focus on Cognitive Neuroscience | Favaloro University

5.5-year study plan, equivalent to Bachelor + Master's degree

- Grade: 9.48/10. Honours Degree, 1st in class.

- Thesis grade: 10/10

2012 – 2017
Argentina

RESEARCH EXPERIENCE

Researcher | CVAI Lab & Ernst Strüngmann Institute (in cooperation with Max-Planck Society)

Studying how AI systems abstract semantic knowledge from unimodal and multimodal sources of information.

2021 – present
Germany

Researcher | Max-Planck-Institute AE

Studied the temporal dynamics and format of neural representations underlying schema-retrieval, episodic-memory, and predictive processing mechanisms, using machine learning methods and representational similarity analysis.

2018 – 2021
Germany

Researcher | COCUCO Lab, Physics Department, University of Buenos Aires

Quantified brain states of reduced consciousness (e.g. anesthesia, sleep) with machine learning methods.

2017 – 2018
Argentina

Intern | LPEN, Institute of Cognitive and Translational Neuroscience (INCyT)

Investigated the neural dynamics of bilingualism with time-frequency analysis.

2014 – 2016
Argentina

ACADEMIC PUBLICATIONS (selected)

(* denotes equal contribution)

AI research

M.G. Vilas, F. Adolfi, D. Poeppel and G. Roig (2024). Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience. *41st International Conference on Machine Learning (ICML)*.

F. Adolfi, **M.G. Vilas**, T. Wareham (2024). The Computational Complexity of Circuit Discovery for Inner Interpretabil-

ity. <https://arxiv.org/pdf/2410.08025>.

F. Adolfi, **M.G. Vilas**, T. Wareham (2024). Complexity-Theoretic Limits on the Promises of Artificial Neural Network Reverse-Engineering. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46.

M.G. Vilas, T. Schaumlöffel and G. Roig (2024). Analyzing vision transformers for image classification in class embedding space. *37th Conference on Neural Information Processing Systems (NeurIPS)*.

M. Prasad Panda, M. Tiezzi, **M.G. Vilas**, G. Roig, B. M. Eskofier, and D. Zanca (2024). FovEx: Human-inspired Explanations for Vision Transformers and Convolutional Neural Networks. *HCV Workshop, ECCV 2024*.

T. Schaumlöffel, **M.G. Vilas** and G. Roig (2023). Peacs: Prefix encoding for auditory caption synthesis. *DCASE2023 Challenge*.

NeuroAI research

M. Guo, B. Choksi, S. Sadiya, A.T. Gifford, **M.G. Vilas**, R.M. Cichy, and G. Roig (2024). Limited but consistent gains in adversarial robustness by co-training object recognition models with human EEG. *HCV Workshop, ECCV 2024*.

A.T. Gifford, B. Lahner, S. Saba-Sadiya, **M.G. Vilas**, A. Lascelles, A. Oliva, K. Kay, G. Roig and R.M. Cichy (2023). The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*.

D. Bersch, K. Dwivedi, **M. Vilas**, R. M. Cichy, and G. Roig (2022). Net2Brain: A Toolbox to compare artificial vision models with human brain responses. *Conference on Cognitive Computational Neuroscience*.

Cognitive Neuroscience research

M.G. Vilas, R. Auksztulewicz, L. Melloni (2021). Active Inference as a Computational Framework for Consciousness. *Review of Philosophy and Psychology*, 1-20.

M.G. Vilas, L. Melloni (2020). A challenge for predictive coding: Representational or experiential diversity? *Behavioral and Brain Sciences*, 43.

M.G. Vilas, L. Melloni (2019). Schema- and episodic-based predictions during visual narrative perception. *The Predictive Brain Conference*, Marseille, France.

C. Pallavacini*, **M.G. Vilas***, M. Villarreal, F. Zamberlan, S. Muthukumaraswamy, D. Nutt, R. Carhart-Harris, E. Tagliazucchi (2019). Spectral signatures of serotonergic psychedelics and glutamatergic dissociatives. *NeuroImage*, 200, 281-291.

M.G. Vilas, M. Santilli, E. Mikulan, F. Adolfi, M. Martorell Caro, F. Manes, E. Herrera, L. Sedeño, A. Ibáñez, A. M. García (2019). Shakespearean tropes and the non-native reader: Age of L2 acquisition modulates neural responses to functional shifts. *Neuropsychologia*, 124, 79-86.

F. Cavanna*, **M.G. Vilas***, M. Palmucci*, E. Tagliazucchi (2018). Dynamic functional connectivity and brain metastability during altered states of consciousness. *NeuroImage*, 180, 383-395.

M. Santilli*, **M.G. Vilas***, E. Mikulan, M. Martorell Caro, E. Muñoz, L. Sedeño, A. Ibáñez, A.M. García (2018). Bilingual memory, to the extreme: Lexical processing in simultaneous interpreters. *Bilingualism: Language and Cognition*, 1-18.

TALKS & TUTORIALS (selected)

M.G. Vilas (2024). Probing the representations and capacities of Vision-Language Models. *Cohere for AI*.

M.G. Vilas (2024). AI Inner Interpretability research through a Cognitive (neuro)Science perspective. Presented

at Max-Planck-Institute for Software Systems, and TU Darmstadt.

S. Saba-Sadiya, **M.G. Vilas**, A. Gifford (2023). Algonauts & Net2Brain Hackathon. *CNN 2023*, Oxford.

M.G. Vilas (2023). Net2Brain: A toolbox to compare artificial deep neural networks with human brain responses. *Data Science Week 2023*, Frankfurt am Main.

M.G. Vilas (2021). Introduction to machine learning and data visualization with Python. *OHBM BrainHack*, online.

M.G. Vilas (2021 & 2020). Computational reproducibility: Best practices outlined by The Turing Way. Presented at *University College London, University of Leicester, EMBL, and Brainhack Donostia*.

M.G. Vilas (2021). Evaluating the reproducibility of deep learning research in cognitive computational neuroscience. *LXAI Social at ICLR 2021*, online.

M.G. Vilas, S. Henin, C. Ranganath, L. Melloni (2021). Schema- and episodic-based predictions during visual narrative perception. *CNS 2021*, online.

M.G. Vilas, K. Whitaker (2021). Why you need a reproducible computational environment and how Binder can help. *Boost your Research Reproducibility with Binder Workshop at 3rd SSI Research Software Camp*, online.

M.G. Vilas, M. Sharan, K. Whitaker (2020). The Turing Way: A guide to reproducible, ethical and collaborative research practices. *LiveMEEG*, online.

HONORS & AWARDS

Open Science SIG Fellowship <i>Organization for Human Brain Mapping (OHBM)</i>	2021
Travel Grant <i>EuroScipy</i>	2019
Ph.D. Scholarship <i>National Scientific and Technical Research Council (CONICET)</i>	2017
Academic Excellence Scholarship <i>Favaloro University</i>	2016
Academic Merit Award <i>Santander Rio Bank</i>	2016, 2014 & 2013

MENTORING

Google Summer of Code <i>Project Mentor</i>	2021
Outreachy <i>Project Mentor</i>	2021
Open Life Science Program <i>Mentor & Expert</i>	2020 & 2021
Book Dash of The Turing Way <i>Mentor / Helper</i>	2020

TEACHING

Guest Lecturer <i>Computer Vision Seminar</i> Goethe University	2024 - pres.
Teaching Assistant <i>Introduction to Machine Learning with scikit-learn</i> Hackathon - Organization for Human Brain Mapping	2021
Instructor <i>Creating a Jupyter Book with The Turing Way</i> JupyterCon 2020	2020
Teaching Assistant <i>Experimental Psychology</i> Favaloro University	2014

OPEN-SCIENCE/OPEN-SOURCE CONTRIBUTIONS

Community Lead - ML Theory <i>Cohere for AI</i>	2024 - pres.
Core Developer <i>net2brain</i>	2021 - pres.
Open Source Contributor <i>scikit-learn, sktime, pandas, jupyter-book, net2brain</i>	2019 - pres.

Core Developer <i>The Turing Way</i>	2020 - 2021
Project Lead <i>Open Life Science Program</i>	2021
Co-organizer <i>pandanistas</i>	2020

SERVICES

Academic

Co-Chair Minisymposium on Neuroscience and Biology <i>SciPy 2021 Conference</i>	2021
Volunteer <i>EuroSciPy 2019 Conference</i>	2019
Reviewer <i>ACL, EMNLP, CVPR, Nature Reviews Neuroscience, Journal of Open Source Software, Current Biology, Frontiers in Human Neuroscience, among others</i>	-

Community

Code of Conduct Committee Member <i>sktime Python Software Package</i>	2020 – 2022
PhD representative <i>Max Planck Institute AE</i>	2019 – 2021