Julien Colin

ELLIS PhD Student

PhD Student
ELLIS Alicante
(+33) 6 42 04 71 00
julien_colin@brown.edu

Education

2022–present **PhD, Computer Science**, *ELLIS Alicante Fundation, University of Alicante*, Spain.
Human-centered Interpretability, under the supervision of Nuria Oliver, Thomas Serre and Miguel Angel Lozano

2020–2021 : **Master 2 Cognitive Sciences, Natural and Artificial Cognition**, *Grenoble Institute of Technology (INP), France.*

2019–2020 : **Master 1 Theorithical and Applied Physics, Biophysics** , *Paul Sabatier University, France.*

2014–2019 : **Bachelor of Science: Physics and Chemistry**, *University of Lorraine, France.*

Research Experience

Brown University, USA

April,2025 – Oct 2025 **Visiting Research Fellow**.
Human alignment and interpretability in visual representations.

Brown University, USA

April,2022 – Sept 2022 **Research Associate**.
Development of methods and metrics for Explainable Artificial Intelligence.

ANITI, France

Sept,2021 – Feb 2022 **Research Associate**.
Development of methods and metrics for Explainable Artificial Intelligence.

ANITI, France

Feb,2021 – July 2021 **Research Intern**.
Development of an evaluation framework for Explainable Artificial Intelligence methods.

Animal Cognition Research Center (CRCA), France

May,2020 – June 2020 **Research Intern**.
Modeling of the movement of bumblebees through Lattice Boltzmann methods.

French National Centre for Scientific Research (CNRS), France

April,2019 – May 2019 **Research Intern**.
Coarse-grained simulation of DNA-proteins repair complex in the cellular environment.

Publications

Work in Progress

2026 **Colin, J.**, Goetschalckx, L., Fel, T., Boutin, V., Serre, T., Oliver, N., Choosing the right basis for interpretability: Psychophysical comparison between neuron-based and dictionary-based representations., *Under review at TMLR.*

2026 **Colin, J.**, Oliver, N., Serre, T., Can human perceptual similarity alignment improve interpretability in visual representations?, *Under review*.

2026 **Colin, J.**, Goetschalckx, L., Oliver, N., Serre, T., Capability $\neq$ Interpretability: Human Interpretability of Vision Foundation Models., *Under review*.

In Conference Proceedings

2026 Colin, J., Oliver, N., and Serre, T. , Does human-alignment benefit interpretability?, Workshop on Representational Alignment (Re^4-Align)(**ICLR26**).

2023 Fel, T., Boissin, T., Boutin, V., Picard, A., Novello, P., **Colin, J.**, Linsley, D., Rousseau, T., Cadène, R., Gardes, L., and Serre, T., Unlocking Feature Visualization for Deeper Networks with MAgnitude Constrained Optimization., Neural Information Processing Systems (**NeurIPS23**).

2023 Boutin, V., Fel, T., Singhal, L., Mukherji, R., Nagaraj, A., **Colin, J.**, and Serre, T., Diffusion models as artists: Are we closing the gap between humans and machines?, International Conference on Machine Learning (**ICML23**).

2023 Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., **Colin, J.**, Cadène, R., and Serre, T. , CRAFT: Concept Recursive Activation FacTorization for explainability, IEEE Conference on Computer Vision and Pattern Recognition (**CVPR23**).

2022 **Colin, J.**, Fel, T., Cadène, R., and Serre, T., What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods., Neural Information Processing Systems (**NeurIPS22**).

2022 Zerroug, A., Vaishnav, M., **Colin, J.**, Musslick, S., and Serre, T. , A benchmark for compositional visual reasoning., Neural Information Processing Systems. Dataset and Benchmarks (**NeurIPS22**).

2022 Fel, T., Hervier, L., Vigouroux, D., Poche, A., Plakoo, J., Cadène, R., Chalvidal, M., **Colin, J.**, Boissin, T., Bethune, L., Picard, A., Nicodeme, C., Gardes, L., Flandin, G., and Serre, T., Xplique: A deep learning explainability toolbox., IEEE Conference on Computer Vision and Pattern Recognition. Workshop on XAI4CV: Explainable Artificial Intelligence for Computer Vision (**CVPR22**).

Journal Article

2024 Riccio, P., **Colin, J.**, Ogolla, S., Oliver, N., Mirror, Mirror on the Wall, Who Is the Whitest of All? Racial Biases in Social Media Beauty Filters, In ***Social Media + Society***.

Scientific Service Community

2022-25 **Co-chair**, ELLIS Reading Group on Human-Centric Machine Learning.

2023-24-25-26 **Reviewer**, NeurIPS.

2024-25-26 **Reviewer**, ICLR: workshop Re-Align.

2026 **Reviewer**, CCN.

2025 **Reviewer**, AAAI: AI Alignement track.

2024 **Reviewer**, NeurIPS: workshop Behavioral ML.

2024 **Reviewer**, ECCV: workshop XAI4CV.

2023 **Reviewer**, CVPR: workshop XAI4CV.

2024 **Reviewer**, ICLR.

2024 **Reviewer**, ICML.

2023-24 **Reviewer**, ICCV/ECCV.

2023 **Program committee and Reviewer**, ECAI.

2022 **Reviewer**, CVPR.

Computer skills

Programming Languages: Python (PyTorch), Javascript