Curriculum Vitae **Andrew M. Saxe**

Personal Details

Andrew M. Saxe
Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre
University College London
25 Howland St Room 484
London W1T 4JG
+44 7541866812
a.saxe@ucl.ac.uk
Nationality: American, British

Current Position

Professorial Research Fellow
Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre
University College London

CIFAR Azrieli Global Scholar, CIFAR
Visiting Professor, University of the Witwatersrand, South Africa

Education

2015	PhD in Electrical Engineering, Stanford University
2010	MS in Electrical Engineering, Stanford University
2008	BSE in Electrical Engineering, Princeton University, *summa cum laude* Concentrations (minors): Robotics & Intelligent Systems; Applications of Computing; Applied and Computational Mathematics

Professional Appointments

2023-	*Professorial Research Fellow*, Gatsby Unit & Sainsbury Wellcome Centre, University College London
2021-2023	*Joint Group Leader*, Gatsby Unit & Sainsbury Wellcome Centre, University College London
2021-2022	*Visiting Researcher*, Meta AI
2020-2021	*Associate Professor*, Department of Experimental Psychology, Oxford University
2019-2021	*Sir Henry Dale Fellow*, Department of Experimental Psychology, Oxford University
2018-2019	*Postdoctoral Research Associate*, Department of Experimental Psychology, Oxford University
2015-2018	*Swartz Postdoctoral Fellow*, Center for Brain Science, Harvard University

Awards and Honors

2023	Blavatnik Awards for Young Scientists in the UK Finalist in Life Sciences, NYAS
2019	Wellcome-Beit Prize, Wellcome Trust
2016	Robert J. Glushko Outstanding Doctoral Dissertations Prize, Cognitive Science Society

2013-2015	Center for Mind, Brain, and Computation Traineeship, Stanford University
2013	Artificial Intelligence Journal Travel Award, CogSci2013
2010-2013	National Defense Science and Engineering Graduate (NDSEG) Fellowship
2010	NSF Graduate Research Fellowship Honorable Mention
2008-2010	Stanford Graduate Fellowship, Stanford University
2008	Hertz Fellowship Finalist
2008	Lore von Jaskowsky Memorial Prize for Contributions to Research, Princeton University
2008	G. David Forney Jr. Prize in Signals & Systems, Princeton University
2007-8	Barry M. Goldwater Scholarship

Grants and Fellowships

2022-2026	Schmidt Science Polymath Award, Schmidt Futures, $2,500,000
2020-2022	CIFAR Azrieli Global Scholarship, CIFAR, CAD100,000
2019-2024	Sir Henry Dale Fellowship, Wellcome Trust and Royal Society, £771,226

Preprints

Carrasco-Davis, Rodrigo, Javier Masís, and Andrew M. Saxe (2023). *Meta-Learning Strategies through Value Maximization in Neural Networks.* arXiv:2310.19919 [cs, q-bio].

Löwe, Anika T., Léo Touzo, Paul S. Muhle-Karbe, Andrew M. Saxe, Christopher Summerfield, and Nicolas W. Schuck (2023). *Regularised neural networks mimic human insight.* arXiv:2302.11351 [cs, q-bio].

Bansal, Y., M. Advani, D.D. Cox, and A.M. Saxe (2018). “Minnorm training: an algorithm for training over-parameterized deep neural networks”. In: *arXiv:1806.00730.*

Publications

Flesch, Timo, David G. Nagy, Andrew Saxe, and Christopher Summerfield (2023). “Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals”. In: *PLOS Computational Biology* 19.1. Publisher: Public Library of Science.

Flesch, Timo, Andrew Saxe, and Christopher Summerfield (Mar. 2023). “Continual task learning in natural and artificial agents”. In: *Trends in Neurosciences* 46.3. Publisher: Elsevier, pp. 199–210.

Jarvis, D., R. Klein, B. Rosman, and A.M. Saxe (2023). “On The Specialization of Neural Modules”. In: *The Eleventh International Conference on Learning Representations.*

Masís, Javier, Travis Chapman, Juliana Y Rhee, David D Cox, and Andrew M Saxe (2023). “Strategically managing learning during perceptual decision making”. In: *eLife* 12.

Nelli, Stephanie, Lukas Braun, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield (2023). “Neural knowledge assembly in humans and neural networks”. In: *Neuron* 111.9, pp. 1504–1516.

Patel, Nishil, Sebastian Lee, Stefano Sarao Mannelli, Sebastian Goldt, and Andrew M. Saxe (2023). “The RL Perceptron: Dynamics of Policy Learning in High Dimensions”. In: *ICLR 2023 Workshop on Physics for Machine Learning.*

Shamash, Philip, Sebastian Lee, Andrew M. Saxe, and Tiago Branco (2023). “Mice identify subgoal locations through an action-driven mapping process”. In: *Neuron* 111.12, pp. 1966–1978.

Singh, Aaditya K., Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill (2023). “The Transient Nature of Emergent In-Context Learning in Transformers”. In: *Thirty-seventh Conference on Neural Information Processing Systems.*

Singh, Aaditya K., David Ding, Andrew Saxe, Felix Hill, and Andrew Kyle Lampinen (2023). “Know your audience: specializing grounded language models with listener subtraction”. In: *17th Conference of the European Chapter of the Association for Computational Linguistics.*

Sun, Weinan, Madhu Advani, Nelson Spruston, Andrew Saxe, and James E. Fitzgerald (2023). “Organizing memories for generalization in complementary learning systems”. In: *Nature Neuroscience* 26.8, pp. 1438–1448.

Braun, Lukas, Clémentine Dominé, James Fitzgerald, and Andrew Saxe (2022). “Exact learning dynamics of deep linear networks with prior knowledge”. In: *Advances in Neural Information Processing Systems.* Vol. 35, pp. 6615–6629.

Flesch, Timo, K. Juechems, T. Dumbalska, A. Saxe*, and Christopher Summerfield* (2022). “Orthogonal representations for robust context-dependent task performance in brains and neural networks”. In: *Neuron* 110. Publisher: Elsevier, *Equal contributions.

Gerace, Federica, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová (2022). “Probing transfer learning with a model of synthetic correlated datasets”. In: *Machine Learning: Science and Technology* 3.1. Publisher: IOP Publishing.

Lee, S., S.S. Mannelli, C. Clopath, S. Goldt, and A.M. Saxe (2022). “Maslow’s Hammer for Catastrophic Forgetting: Node Re-Use vs Node Activation”. In: *ICML.*

Saglietti, Luca, Stefano Mannelli, and Andrew Saxe (2022). “An Analytical Theory of Curriculum Learning in Teacher-Student Networks”. In: *Advances in Neural Information Processing Systems.* Vol. 35, pp. 21113–21127.

Saxe, Andrew, Shagun Sodhani, and Sam Jay Lewallen (2022). “The Neural Race Reduction: Dynamics of Abstraction in Gated Networks”. In: *Proceedings of the 39th International Conference on Machine Learning.* ISSN: 2640-3498. PMLR, pp. 19287–19309.

Juechems, K. and A. Saxe (2021). “Inferring Actions, Intentions, and Causal Relations in a Deep Neural Network”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 43.

Lee, S., S. Goldt, and A. Saxe (2021). “Continual Learning in the Teacher-Student Setup: Impact of Task Similarity”. In: *Proceedings of the 38th International Conference on Machine Learning.*

Saxe, A., S. Nelli, and C. Summerfield (2021). “If deep learning is the answer, what is the question?” In: *Nature Reviews Neuroscience* 22.1, pp. 55–67.

Advani*, M.S., A.M. Saxe*, and H. Sompolinsky (2020). “High-dimensional dynamics of generalization error in neural networks”. In: *Neural Networks* 132, 428–446. *Equal contribution.

Cao, Y., C. Summerfield, and A. Saxe (2020). “Characterizing emergent representations in a space of candidate learning rules for deep networks”. In: *Advances in Neural Information Processing Systems 33.*

Goldt, S., M.S. Advani, A.M. Saxe, F. Krzakala, and L. Zdeborová (2020). “Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.12. Publisher: IOP Publishing, p. 124010.

Musslick, S., A. Saxe, A.N. Hoskin, D. Reichman, and J.D. Cohen (2020). *On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations.* Tech. rep. type: article. PsyArXiv.

Goldt, S., M.S. Advani, A.M. Saxe, F. Krzakala, and L. Zdeborová (2019). “Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup”. In: *NeurIPS.* arXiv: 1906.08632. Oral presentation.

Richards, B.A., T.P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann

Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording (2019). “A deep learning framework for neuroscience”. In: *Nature Neuroscience* 22.11, pp. 1761–1770.

Saxe, A.M., Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B.D. Tracey, and D.D. Cox (Dec. 2019). “On the information bottleneck theory of deep learning”. In: *Journal of Statistical Mechanics: Theory and Experiment* 12. Publisher: IOP Publishing, p. 124020.

Saxe, A.M., J.L. McClelland, and S. Ganguli (2019). “A mathematical theory of semantic development in deep neural networks”. In: *Proceedings of the National Academy of Sciences* 116.23. arXiv: 1810.10531, pp. 11537–11546.

Earle, A.C., A.M. Saxe, and B. Rosman (2018). “Hierarchical Subtask Discovery with Non-Negative Matrix Factorization”. In: *International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. Vancouver, Canada.

Nye, M. and A. Saxe (2018). “Are Efficient Deep Representations Learnable?” In: *Workshop Track at the International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. arXiv: 1511.06434v1 ISSN: 0004-6361. Vancouver, Canada.

Saxe, A.M., Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B.D. Tracey, and D.D. Cox (2018). “On the Information Bottleneck Theory of Deep Learning”. In: *International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. Vancouver, Canada.

Zhang, Y., A.M. Saxe, M.S. Advani, and A.A. Lee (2018). “Energy-entropy competition and the effectiveness of stochastic gradient descent in machine learning”. In: *Molecular Physics*. arXiv: 1803.01927, pp. 1–10.

Earle, A.C., A.M. Saxe, and B. Rosman (2017). “Hierarchical Subtask Discovery With Non-Negative Matrix Factorization”. In: *Workshop on Lifelong Learning: A Reinforcement Learning Approach at ICML*. arXiv: 1708.00463v1 Place: Sydney, Australia.

Musslick, S., A.M. Saxe, K. Ozcimder, B. Dey, G. Henselman, and J.D. Cohen (2017). “Multitasking Capability Versus Learning Efficiency in Neural Network Architectures”. In: *Annual meeting of the Cognitive Science Society*, pp. 829–834.

Saxe, A.M., A.C. Earle, and B. Rosman (2017). “Hierarchy Through Composition with Multitask LMDPs”. In: *International Conference on Machine Learning*. Sydney, Australia.

McClelland, J.L., Z. Sadeghi, and A.M. Saxe (2016). “A Critique of Pure Hierarchy: Uncovering Cross-Cutting Structure in a Natural Dataset”. In: *Neurocomputational Models of Cognitive Development and Processing*. Publisher: World Scientific, pp. 51–68.

Tsai*, C.Y., A. Saxe*, and D. Cox (2016). “Tensor Switching Networks”. In: *Advances in Neural Information Processing Systems 29*. arXiv: 1610.10087 ISSN: 10495258. *Equal contributions.

Goodfellow, I.J., O. Vinyals, and A.M. Saxe (2015). “Qualitatively Characterizing Neural Network Optimization Problems”. In: *International Conference on Learning Representations*. arXiv: 1412.6544v4. San Diego, CA: Oral presentation.

Saxe, A.M., J.L. McClelland, and S. Ganguli (2014). “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks”. In: *International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. arXiv: 1312.6120v3. Banff, Canada: Oral presentation.

Saxe, A.M., J.L. McClelland, and S. Ganguli (2013b). “Dynamics of learning in deep linear neural networks”. In: *NIPS Workshop on Deep Learning*.

Saxe, Andrew M., James L. McClelland, and Surya Ganguli (2013). “Learning hierarchical category structure in deep neural networks”. In: *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Balci, F., P. Simen, R. Niyogi, A. Saxe, J.A. Hughes, P. Holmes, and J.D. Cohen (2011). “Acquisition of decision making criteria: reward rate ultimately beats accuracy”. In: *Attention, Perception, & Psychophysics* 73.2. Publisher: Springer, pp. 640–57.

Saxe, A., M. Bhand, R. Mudur, B. Suresh, and A.Y. Ng (2011). “Unsupervised learning models of primary cortical receptive fields and receptive field plasticity”. In: *Advances in Neural Information Processing Systems 25*.

Saxe, A.M., P.W. Koh, Z. Chen, M. Bhand, B. Suresh, and A.Y. Ng (2010). “On Random Weights and Unsupervised Feature Learning”. In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Baldassano, C.A., G.H. Franken, J.R. Mayer, A.M. Saxe, and D.D. Yu (2009). “Kratos: Princeton University’s entry in the 2008 Intelligent Ground Vehicle Competition”. In: *Proceedings of SPIE*.

Goodfellow, I.J., Q.V. Le, A.M. Saxe, H. Lee, and A.Y. Ng (2009). “Measuring Invariances in Deep Networks”. In: *Advances in Neural Information Processing Systems 24*. Ed. by Y. Bengio and D. Schuurmans.

Atreya, A.R., B.C. Cattle, B.M. Collins, B. Essenburg, G.H. Franken, A.M. Saxe, S.N. Schiffres, and A.L. Kornhauser (2006). “Prospect Eleven: Princeton University’s entry in the 2005 DARPA Grand Challenge”. In: *Journal of Field Robotics* 23.9, pp. 745–753.

Invited Presentations

2023	Workshop on Analytical Approaches for Neural Network Dynamics, Paris
2023	FENS Brain Conference “Structuring knowledge for flexible behaviour”, Copenhagen
2023	Statistical Physics and Machine Learning Back Together Again, Cargese
2023	Workshop on Statistical Learning and the Brain, KITP
2023	Titisee Conference on NeuroAI, Germany
2023	NeuroStatPhys Workshop, Les Houches
2022	CIFAR Workshop on Neurofoundation Models, Montreal
2022	Invited Lecture, Bernstein Computational Neuroscience Conference, Berlin
2022	ELLIS Natural Intelligence Workshop, Crete
2022	Main Lecturer, Lake Como School on Statistical Physics of Deep Learning
2022	The Great AI Debate, FENS, Paris
2022	FENS Symposium on Single Neurons and AI
2022	Harvard ML Foundations Seminar, Harvard
2022	Methods in Computational Neuroscience, Norway
2021	EPFL Applied ML Day, Lausanne
2021	Plenary Lecture, IUPAP Conference on Computational Physics, Coventry
2021	Methods in Computational Neuroscience, Norway
2020	CIFAR Deep Learning Summer School, Montreal
2020	AI and the Brain Symposium, ETH, Zurich
2020	Computational Neuroscience Seminar, TU Berlin
2020	Machine Learning Applications to Physics, Princeton Center for Theoretical Science
2019	Analyses of Deep Learning (STATS 385), Stanford University
2019	Chaucer Club Seminar, Cambridge University
2019	Workshop on Science of Data Science, ICTP, Trieste
2019	Istanbul Workshop on Theory of DL, IMBM
2019	SfN Machine Learning Virtual Conference
2019	ICML Workshop on Deep Learning Phenomena, Long Beach
2019	Mind and Machine Seminar, University of Bristol
2019	Universitat Pompeu Fabra, Barcelona
2019	Bellairs Workshop on Deep Learning and Neuroscience, Barbados
2018	Statistical Physics Seminar, ENS, Paris
2018	PDP Symposium, Princeton
2018	Computation and Theory Seminar, Janelia
2018	Symposium on the Mathematical Theory of Deep Neural Networks, Princeton
2017	Oxford Neurotheory Forum, Oxford
2017	Temporal Dynamics of Learning Seminar, UCSD

2016	Google DeepMind, London
2016	15th Neural Computation and Psychology Workshop, Philadelphia
2016	Google Research, Cambridge, MA
2016	Deep Learning Workshop, Center for Brains, Minds, and Machines, MIT
2016	Redwood Center for Theoretical Neuroscience, UC Berkeley
2016	Apple, Cupertino, CA
2015	Brains, Minds, and Machines Symposium, NIPS, Montreal

Teaching

2021-	Theoretical Neuroscience Core Course, deep learning theory module, Gatsby Unit, UCL
2021-	Systems Neuroscience and Theoretical Neuroscience, cognition module, Sainsbury Wellcome Centre, UCL
2021-	Lecture, Cognitive and Decision Science, UCL
2020-2021	Lecture in Computational Neuroscience module, MSc in Neuroscience, Oxford University
2019	Teaching Assistant, Connectionism Block Practical, Oxford University
2018	Distinction in Teaching Award (NEURO120), Harvard University
2017	Course Designer, Introductory Computational Neuroscience (NEURO120), Harvard University
2017	Distinction in Teaching Award (MCB131), Harvard University
2017	Head Teaching Fellow, MCB131: Computational Neuroscience, Harvard University
2014	Guest Lecturer, PSYCH209: Neural network and deep learning models for cognition and cognitive neuroscience, Stanford University
2010	Teaching Assistant, CS294A: Research projects in Artificial Intelligence, Stanford University
2009	Teaching Assistant, CS229: Machine Learning, Stanford University

Service to Profession

Conference and School Organizer

2023-2024	Workshop Chair, Computational and Systems Neuroscience Conference
2023-	Founding Organizer, Analytical Connectionism Summer School
2023	Program Committee, Bernstein Computational Neuroscience Conference
2019	Conference on the Mathematical Theory of Deep Neural Networks, New York
2019	Conference on Deep Learning and the Brain, Jerusalem, Israel

Workshop Organizer

2020	Panel at Oxford Autumn School in Neuroscience, Virtual
2020	Panel at International Conference on Mathematical Neuroscience, Virtual
2019	Cosyne 2019 Workshop on continual learning in biological and artificial neural networks
2016	CogSci 2016 Tutorial Workshop on Contemporary Deep Neural Network Models, Philadelphia
2014	CogSci 2014 Workshop on Deep Learning and the Brain, Quebec City, Cananda

Journal Reviewer

Nature Communications
Proceedings of the National Academy of Sciences (PNAS)
Journal of Machine Learning Research (JMLR)
PLOS ONE

Neural Computation
IEEE Transactions on Neural Networks and Learning Systems (IEEE-TNNLS)
IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE-TPAMI)
IEEE Transactions on Knowledge and Data Engineering (IEEE-TKDE)

Conference Reviewer

International Conference on Machine Learning (ICML)
Advances in Neural Information Processing Systems (NIPS) (Reviewer Award, 2013 and 2017)
International Conference on Learning Representations (ICLR) (Reviewer Award, 2017)
International Conference on Artificial Intelligence and Statistics (AISTATS)
Cognitive Science Society Annual Meeting (CogSci)

University Service

High Performance Computing Committee, 2021-
SWC/GCNU Athena SWAN SAT and EDI Working Group, 2021-
Department of Experimental Psychology Athena SWAN Working Group, 2019-2021
Congregation, University of Oxford, 2020

Community Involvement and Outreach

2023 Mathematics for Psychologists Bridging Program, UCL
2021 Day Lead, Neuromatch Deep Learning Academy

Last updated: November 14, 2023 • Typeset in XƎTEX