

Assistant Professor: University of Montreal
 Department of Computer Science and Operations Research
 Quebec AI Institute (Mila)

November 2024 –

Assistant Professor: University of Cambridge
 Department of Engineering, Information Engineering Division
 Computational and Biological Learning Lab (CBL)

June 2021 – Sept 2024

EDUCATION **University of Montreal** / Mila, Montreal, QC, Canada
 Ph.D., Computer Science, June 2022

- Thesis: *AI Alignment and Generalization in Deep Learning*
 - Supervisor: **Aaron Courville**
 - Committee: Yoshua Bengio, Roger Grosse, Guillaume Lajoie
- M.Sc., Computer Science, Dec 2015
- Thesis: *Designing Regularizers and Architectures for Recurrent Neural Networks*
 - Supervisors: **Yoshua Bengio, Roland Memisevic**

Reed College, Portland, OR, USA

B.A., Mathematics, May 2011

- Thesis: *Extending the Critical Group to Oriented Matroids and Simplicial Complexes*
- Supervisor: **David Perkinson**

Budapest Semesters in Mathematics, Fall 2009.

GRANTS AND AWARDS	<ul style="list-style-type: none"> • CIFAR AI Chair: 1,150,000 CAD • Schmidt Sciences: \$300,000 • IVADO professorship in Responsible AI: \$500,000 CAD • Marshall School Distinguished Young Alumni Award • Open Philanthropy Project: \$1,000,000 • Survival and Flourishing Fund: \$880,000 • Open Philanthropy Project: \$250,000 • Long-Term Future Fund: \$200,000 • 1st place (judges) and 1st place (people's choice): (Canada-wide) AI Can Trainee 3-Minute Impact Competition • Effective Altruism Foundation Fund: \$10,000 • Assisted in sourcing/writing Open Philanthropy Project grant to Mila: \$2,400,000 • NVIDIA Pioneering Research Award • University of Montreal Departmental Excellence Scholarship: \$2,500 • Most OpenReview comments at ICLR 	2025
		2024
		2024
		2023
		2022
		2021
		2021
		2021
		2021
		2019
		2017
		2017
		2016
		2014

OTHER AFFILIATIONS	<ul style="list-style-type: none"> • Research Affiliate: Center for the Study of Existential Risk (CSER) Feb 2022 – present • External Associate Academic Member: Quebec AI Institute (Mila) Sept 2022 – Nov 2024 • Member: European Laboratory for Learning and Intelligent Systems (ELLIS) Sept 2022 – present • Affiliate: Center for Human Compatible AI (CHAI) Sept 2022 – present • Member: Institute for Advanced Study (IAS) AI Policy and Governance Working Group June 2023 – present • Board Member: Center for AI Policy (CAIP) Jan 2024 – June 2025 • Fellow: The Alan Turing Institute March 2024 – present
---------------------------	---

PROFESSIONAL EXPERIENCE	<ul style="list-style-type: none"> • Research Director: UK Frontier AI Task Force September 2023 – December 2023 Assisted in organizing the first global AI Safety Summit. Co-authored “Frontier AI: capabilities and risks – discussion paper”. Supervisor: Oliver Ilott • Research Intern: DeepMind Feb-July 2018 Topic: Managing incentives of AI systems. Supervisor: Jan Leike • Research Intern: ElementAI Sept-Dec 2017 Topic: Normalizing flows and Bayesian deep learning. Supervisor: Alexandre Lacoste • Contract Researcher: Partnership on AI (PAI) Sept-Oct 2017 Synthesize academic and stakeholder views in 10-page safety-critical AI executive primer. • Freelance Career Mentor: 80,000 hours July 2016 - Sept 2017 Provide AI career advice for newcomers interested in AI Alignment, ~2hrs/week. • Research Intern: Future of Humanity Institute, Oxford July-Sept 2016 Topic: Learning human preferences from limited feedback. Supervisor: Owain Evans • Contract Reporter: CIFAR Deep Learning Summer School 2015 • Research Assistant: Baylor College of Medicine June-August 2009 Topic: Preliminary work related to “Optimal Inference of Sameness.” (<i>PNAS</i>, 2011). Supervisors: Wei Ji Ma, Krešimir Josić • Dish Machine Operator: The New Scenic Cafe 2006-2007 Donated majority of earnings (roughly \$5000) to Nothing But Nets.
REFEREED PUBLICATIONS	<ol style="list-style-type: none"> 1. Bruno Mlozozeniec, Isaac Reid, Sam Power, David Krueger, Murat Erdogdu, Richard E Turner, Roger Grosse (2025). Distributional Training Data Attribution. <i>Neural Information Processing Systems (spotlight)</i>. 2. Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Curtis Mozer, Eleni Triantafillou (2025). From Dormant to Deleted: Tamper-Resistant Unlearning Through Weight-Space Regularization. <i>Neural Information Processing Systems</i>. 3. Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, Dmitrii Krasheninnikov (2025). Detecting High-Stakes Interactions with Activation Probes. <i>Neural Information Processing Systems</i>. 4. Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, David Duvenaud (2025). Position: Humanity Faces Existential Risk from Gradual Disempowerment. <i>International Conference on Machine Learning</i>. 5. Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, Joar Max Viktor Skalse (2025). The Perils of Optimizing Learned Reward Functions: Low Training Error Does Not Guarantee Low Regret. <i>International Conference on Machine Learning</i>.

⁰The † symbol indicates equal contribution.

6. Tingchen Fu, Mrinank Sharma, Philip Torr, Shay B Cohen, **David Krueger**, Fazl Barez (2025). *PoisonBench: Assessing Large Language Model Vulnerability to Data Poisoning*. *International Conference on Machine Learning*.
7. Bruno Kacper Mlodzeniec, **David Krueger**, Richard E Turner (2025). Position: Probabilistic Modelling is Sufficient for Causal Inference. *International Conference on Machine Learning*.
8. Minseon Kim, Jin Myung Kwak, Lama Alssum, Bernard Ghanem, Philip Torr, **David Krueger**, Fazl Barez, Adel Bibi (2025). Rethinking Safety in LLM Fine-tuning: An Optimization Perspective. *Conference on Language Modeling*.
9. Abhinav Menon, Manish Shrivastava, **David Krueger**, Ekdeep Singh Lubana (2025). Analyzing (In) Abilities of SAEs via Formal Languages. *Nations of the Americas Chapter of the Association for Computational Linguistics*.
10. Matthew Farrugia-Roberts, Karim Ahmed Abdel Sadek, Hannah Erlebach, Christian Schroeder de Witt, **David Krueger**, Usman Anwar, Michael D Dennis (2025). Mitigating Goal Misgeneralization via Minimax Regret. *Reinforcement Learning Conference*.
11. Usman Anwar, Johannes von Oswald, Louis Kirsch, **David Krueger**, Spencer Frei (2025). Understanding In-Context Learning of Linear Models in Transformers Through an Adversarial Lens. *Transactions on Machine Learning Research*.
12. Jan Wehner, Sahar Abdelnabi, Daniel Tan, **David Krueger**, Mario Fritz (2025). Taxonomy, opportunities, and challenges of representation engineering for large language models. *Transactions on Machine Learning Research*.
13. Shoaib Ahmed Siddiqui, Radhika Gaonkar, Boris Köpf, **David Krueger**, Andrew Paverd, Ahmed Salem, Shruti Tople, Lukas Wutschitz, Menglin Xia, Santiago Zanella-Béguelin (2025). Permissive Information-Flow Analysis for Large Language Models. *Transactions on Machine Learning Research*.
14. Neel Alex, Shoaib Ahmed Siddiqui, Amartya Sanyal, **David Krueger** (2025). Protecting Against Simultaneous Data Poisoning Attacks. *International Conference on Learning Representations*.
15. Stephen Casper, David Krueger, Dylan Hadfield-Menell (2025). Pitfalls of Evidence-Based AI Policy. *International Conference on Learning Representations*.
16. Jakub Vrabel, Ori Shem-Ur, Yaron Oz, **David Krueger** (2025). Input Space Mode Connectivity in Deep Neural Networks. *International Conference on Learning Representations*.
17. Bruno Kacper Mlodzeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, **David Krueger**, Richard E. Turner (2025). Influence Functions for Scalable Data Attribution in Diffusion Models. *International Conference on Learning Representations (Oral)*.
18. Clement Neo, Luke Ong, Philip Torr, Mor Geva, **David Krueger**, Fazl Barez (2025). Towards Interpreting Visual Information Processing in Vision-Language Models. *International Conference on Learning Representations*.
19. Thomas Bush, Stephen Chung, Usman Anwar, Adrià Garriga-Alonso, **David Krueger** (2025). Interpreting Emergent Planning in Model-Free Reinforcement Learning. *International Conference on Learning Representations (Oral)*.
20. Stephen Chung, Scott Niekum, **David Krueger** (2024). Predicting Future Actions of Reinforcement Learning Agents. *Neural Information Processing Systems*.

21. Luke Marks, Amir Abdullah, Clement Neo, Rauno Arike, **David Krueger**, Philip Torr, Fazl Barez (2024). *Interpreting Learned Feedback Patterns in Large Language Models*. *Neural Information Processing Systems*.
22. Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, **David Krueger** (2024). *Stress-Testing Capability Elicitation With Password-Locked Models*. *Neural Information Processing Systems*
23. James Urquhart Allingham, Bruno Kacper Mlodzeniec, Shreyas Padhy, Javier Antorán, **David Krueger**, Richard E. Turner, Eric Nalisnick, Jose Miguel Hernández-Lobató (2024). *A Generative Model of Symmetry Transformations*. *Neural Information Processing Systems*.
24. Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigearthaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, **David Krueger** (2024). *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*. *Transactions on Machine Learning Research*.
25. Dmitrii Krasheninnikov, Egor Krasheninnikov, Bruno Mlodzeniec, Tegan Maharaj, **David Krueger** (2024). *Implicit meta-learning may lead language models to trust more reliable sources*. *International Conference on Machine Learning*.
26. Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, **David Krueger**, Noam Kolt, Lennart Heim, Markus Anderljung (2024). *Visibility into AI Agents*. *ACM Conference on Fairness, Accountability, and Transparency*.
27. Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémie Scheurer, Marius Hobbahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, **David Krueger**, Dylan Hadfield-Menell (2024). *Black-Box Access is Insufficient for Rigorous AI Audits*. *ACM Conference on Fairness, Accountability, and Transparency*.
28. Shoaib Ahmed Siddiqui, **David Krueger**, Yann LeCun, Stephane Deny (2024). *Blockwise Self-Supervised Learning at Scale*. *Transactions on Machine Learning Research*
29. Samyak Jain[†], Robert Kirk[†], Ekdeep Singh Lubana[†], Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, **David Krueger** (2024). *Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks*. *International Conference on Learning Representations*.
30. Thomas Coste, Usman Anwar, Robert Kirk, **David Krueger** (2024). *Reward Model Ensembles Help Mitigate Overoptimization*. *International Conference on Learning Representations*.
31. Cindy Wu, Ekdeep Singh Lubana, Bruno Mlodzeniec, Robert Kirk, **David Krueger** (2024). *What Mechanisms Does Knowledge Distillation Distill? UniReps: the First Workshop on Unifying Representations in Neural Models.*¹

¹This work was accepted as an archival submission to the workshop.

32. Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atilim Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, **David Krueger**, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, Sören Mindermann (2023). Managing AI Risks in an Era of Rapid Progress. *Nature*.
33. Bruno Mloedeniec, **David Krueger**, Richard Turner (2024). Implicitly Bayesian Prediction Rules in Deep Learning. *Proceedings of the Symposium on Advances in Approximate Bayesian Inference*.
34. Stephen Casper[†], Xander Davies[†], Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, **David Krueger**, Dorsa Sadigh, Dylan Hadfield-Menell (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.
35. Stephen Chung, Ivan Anokhin, **David Krueger** (2023). Thinker: Learning to Plan and Act. *Neural Information Processing Systems*.
36. Micah Carroll[†], Alan Chan[†], Henry Ashton, **David Krueger** (2023). Characterizing Manipulation from AI Systems. *ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.
37. Alan Chan, Rebecca Salganik, Zhonghao He, John Burden, Yawen Duan, Shalaleh Rismani, Alva Markelius, Katherine Collins, Maryam Molamohammadi, Chris Pang, Lauro Langosco, Konstantinos Voudouris, Wanru Zhao, Dmitrii Krasheninnikov, Michelle Lin, Alex Mayhew, Umang Bhatt, Adrian Weller, **David Krueger**, Tegan Maharaj (2023). Harms from Increasingly Agentic Algorithmic Systems. *ACM Conference on Fairness, Accountability, and Transparency*.
38. Ekdeep Singh Lubana, Eric J Bigelow, Robert Dick, **David Krueger**[†], Hidenori Tanaka[†] (2023). Mechanistic Mode Connectivity. *International Conference on Machine Learning*.
39. Shoaib Ahmed Siddiqui, Nitashan Rajkumar, Tegan Maharaj, **David Krueger**, Sara Hooker (2023). Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics. *International Conference on Learning Representations (spotlight / top 25%)*.
40. Ethan Caballero, Kshitij Gupta, Irina Rish, **David Krueger** (2023). Broken Neural Scaling Laws. *International Conference on Learning Representations*.
41. Joar Skalse[†], Niki Howe, Dmitrii Krasheninnikov, **David Krueger**[†] (2022). Defining and Characterizing Reward Gaming. *Neural Information Processing Systems*.
42. Lauro Langosco Di Langosco[†], Jack Koch[†], Lee Sharkey[†], Jacob Pfau, **David Krueger** (2022). Goal Misgeneralization in Deep Reinforcement Learning. *International Conference on Machine Learning*.
43. **David Krueger**, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinhui Zhang, Rémi Le Priol, Aaron Courville (2021). Out-of-Distribution Generalization via Risk Extrapolation (REx). *International Conference on Machine Learning (Oral)*.

44. Shahar Avin, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, Igor Krawczuk, **David Krueger**, Jonathan Lebensold, Tegan Maharaj, Noa Zilberman (2021). Filling gaps in trustworthy development of AI. *Science*.
45. Chin-Wei Huang[†], **David Krueger**[†], Alexandre Lacoste, Aaron Courville (2018). Neural Autoregressive Flows. *International Conference on Machine Learning*.
46. Joel Moniz, **David Krueger** (2017). Nested LSTMs. *Asian Conference on Machine Learning*.
47. Devansh Arpit[†], Stanislaw Jastrzbski[†], Nicolas Ballas[†], **David Krueger**[†], Emmanuel Bengio, Maxinder Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, Simon Lacoste-Julien (2017). A Closer Look at Memorization in Deep Networks. *International Conference on Machine Learning*.
48. **David Krueger**[†], Tegan Maharaj[†], János Kramár, Mohammad Pezeshki, Nicolas Ballas, Nan Rosemary Ke, Anirudh Goyal, Yoshua Bengio, Aaron Courville, Chris Pal (2017). Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations. *International Conference on Learning Representations*.
49. **David Krueger** and Roland Memisevic (2016). Regularizing RNNs by Stabilizing Activations. *International Conference on Learning Representations (Oral)*.
50. Roland Memisevic, Kishore Konda, **David Krueger** (2015). Zero-bias Autoencoders and the benefits of co-adapting features. *International Conference on Learning Representations*.

OTHER
PUBLICATIONS

1. Dmitrii Krasheninnikov, Richard E Turner, **David Krueger** (2025). Language models' activations linearly encode training-order recency. *Workshop on the Impact of Memorization on Trustworthy Foundation Models at ICML*.
2. Joschka Braun, Carsten Eickhoff, **David Krueger**, Seyed Ali Bahrainian, Dmitrii Krasheninnikov (2025). Understanding (Un) Reliability of Steering Vectors in Language Models. *Building Trust in LLMs and LLM Applications Workshop at ICML*.
3. Usman Anwar, Johannes Von Oswald, Louis Kirsch, **David Krueger**, Spencer Frei (2025). Adversarial Robustness of In-Context Learning in Transformers for Linear Regression.
4. Sina Däubener, Kira Maag, Simon Heilig, **David Krueger**, Asja Fischer (2025). Integrating Uncertainty Quantification into Randomized Smoothing-Based Robustness Guarantees.
5. Brandon Jaipersaud, **David Krueger**, Ekdeep Singh Lubana (2025). How Do LLMs Persuade? Linear Probes Can Uncover Persuasion Dynamics in Multi-Turn Conversations.
6. Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, **David Krueger**, Sören Mindermann, José Hernandez-Orallo, Mor Geva, Yarin Gal (2025). Open Problems in Machine Unlearning for AI Safety.
7. Lauro Langosco, William Baker, Neel Alex, Herbie Bradley, David Quarel, **David Krueger** (2025). Towards Meta-Models for Automated Interpretability.
8. Luke Marks, Alasdair Paren, **David Krueger**, Fazl Barez (2025). Enhancing Neural Network Interpretability with Feature-Aligned Sparse Autoencoders.

9. Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, **David Krueger**, Lennart Heim, Markus Anderljung (2025). IDs for AI Systems.
10. Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, **David Krueger**, Fazl Barez (2025). Quantifying Feature Space Universality Across Large Language Models via Sparse Autoencoders.
11. Jose Miguel Lara Rangel, Stefan Schoepf, Jack Foster, **David Krueger**, Usman Anwar (2024). Learning to forget using hypernetworks. *Workshop on New Frontiers in Adversarial Machine Learning at NeurIPS*.
12. Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, **David Krueger** (2024). Towards Reliable Evaluation of Behavior Steering Interventions in LLMs. *Workshop on Foundation Model Interventions (MINT) at NeurIPS (Oral)*.
13. Akash Wasil, Joshua Clymer, **David Krueger**, Emily Dardaman, Simeon Campos, Evan Murphy (2024). Affirmative Safety: An Approach to Risk Management for Advanced AI.
14. Madeline Brumley, Joe Kwon, **David Krueger**, Dmitrii Krasheninnikov, Usman Anwar (2024). Comparing Bottom-Up and Top-Down Steering Approaches on In-Context Learning Tasks. *Workshop on Foundation Model Interventions (MINT) at NeurIPS*.
15. Dmitrii Krasheninnikov, **David Krueger** (2024). Steering Clear: A Systematic Study of Activation Steering in a Toy Setup. *Workshop on Foundation Model Interventions (MINT) at NeurIPS*.
16. Joshua Clymer, Nick Gabrieli, **David Krueger**, Thomas Larsen (2024). Safety Cases: How to Justify the Safety of Advanced AI Systems.
17. Shoaib Ahmed Siddiqui, Xin Dong, Greg Heinrich, Thomas Breuel, Jan Kautz, **David Krueger**, Pavlo Molchanov (2024). A deeper look at depth pruning of LLMs. *Workshop on Theoretical Foundations of Foundation Models at ICML*.
18. Lauro Langasco, Neel Alex, William Baker, David Quarel, Herbie Bradley, **David Krueger** (2024). Detecting Backdoors with Meta-Models. *Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly at NeurIPS*.
19. Shoaib Ahmed Siddiqui, Jean Kossaifi, Boris Bonev, Christopher Choy, Jan Kautz, **David Krueger**, Kamyar Azizzadenesheli (2024). Exploring the Design Space of Deep-Learning-Based Weather Forecasting Systems.
20. Diego Dorn, Neel Alex, **David Krueger** (2023). Goal Misgeneralization as Implicit Goal Conditioning. *Workshop on Goal-Conditioned Reinforcement Learning at NeurIPS*.
21. Shoaib Ahmed Siddiqui, **David Krueger**, Thomas Breuel (2023). Investigating the Nature of 3D Generalization in Deep Neural Networks.
22. William Baker, Herbie Bradley, **David Krueger** (2023). Inverse Tracr: Mapping Neural Network Weights to Code.
23. Alan Chan[†], Ben Bucknall[†], Herbie Bradley, **David Krueger** (2024). Hazards from Increasingly Accessible Fine-Tuning of Downloadable Foundation Models.
24. Samuel Curtis, Ravi Iyer, Cameron Domenico Kirk-Giannini, Victoria Krakovna, **David Krueger**, Nathan Lambert, Bruno Marnette, Colleen McKenzie, Julian Michael, Evan Miyazono, Noyuri Mima, Aviv Ovadya, Luke Thorburn, Deger Turan (2024). Research Agenda for Sociotechnical Approaches to AI Safety

25. Lauro Langosco, **David Krueger**, Adam Gleave (2023). Training Equilibria in Reinforcement Learning.
26. Adam Ibrahim, Charles Guille-Escuret, Ioannis Mitliagkas, Irina Rish, **David Krueger**, Pouya Bashivan (2023). Towards Out-of-Distribution Adversarial Robustness.
27. Usman Anwar, Jia Wan, **David Krueger**, Jakob Foerster (2023). Noisy ZSC: Breaking The Common Knowledge Assumption In Zero-Shot Coordination Games. *Second Agent Learning in Open-Endedness Workshop at NeurIPS*.
28. Xander Davies, Lauro Langosco, **David Krueger** (2022). Unifying Grokking and Double Descent. *Machine Learning Safety NeurIPS workshop*.
29. Lev McKinney, Yawen Duan, **David Krueger**, Adam Gleave (2022). On The Fragility of Learned Reward Functions. *Machine Learning Safety NeurIPS workshop*.
30. Alan Clark, Shoaib Ahmed Siddiqui, Robert Kirk, Usman Anwar, Stephen Chung, **David Krueger** (2022). Domain Generalization for Robust Model-Based Offline RL. *Offline RL NeurIPS workshop*.
31. Dmitrii Krasheninnikov, Egor Krasheninnikov, **David Krueger** (2022). Assistance with large language models. *Machine Learning Safety NeurIPS workshop*.
32. Enoch Tetteh, Joseph Viviano, Yoshua Bengio, **David Krueger**, Joseph Paul Cohen (2021). Multi-Domain Balanced Sampling Improves Out-of-Distribution Generalization of Chest X-ray Pathology Prediction Models. *Medical Imaging Meets NeurIPS workshop*.
33. Andrew Critch and **David Krueger** (2020). AI Research Considerations for Human Existential Safety (ARCHEs).
34. Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger [and 54 others, including **David Krueger**] (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.
35. **David Krueger**, Tegan Maharaj, Jan Leike (2020). Hidden Incentives for Auto-induced Distributional Shift.
36. **David Krueger**, Tegan Maharaj, Shane Legg, Jan Leike (2019). Misleading Meta-objectives and Hidden Incentives for Distributional Shift. *ICLR workshop on Safe Machine Learning (Oral)*.
37. Jan Leike, **David Krueger**, Tom Everitt, Miljan Martic, Vishal Maini, Shane Legg (2018). Scalable Agent Alignment via Reward Modeling: a Research Direction.
38. **David Krueger**[†], Chin-Wei Huang[†], Riahsat Islam, Ryan Turner, Alexandre Lacoste, Aaron Courville (2017). Bayesian Hypernetworks. *NeurIPS workshop on Bayesian Deep Learning*.
39. Alexandre Lacoste, Thomas Boquet, Negar Rostamzadeh, Boris Oreshkin, Wonchang Chung, **David Krueger** (2017). Deep prior. *NeurIPS workshop on Bayesian Deep Learning*.
40. **David Krueger**, Jan Leike, John Salvatier, Owain Evans (2016). Active Reinforcement Learning: Observing Rewards at a Cost. *NeurIPS workshop on The Future of Interactive Learning Machines (FILM)*.
41. Laurent Dinh, **David Krueger**, Yoshua Bengio (2015). NICE: Nonlinear Independent Component Estimation. *ICLR workshop*.

TEACHING	<ul style="list-style-type: none"> • University of Montreal: AI Safety and Alignment • Cambridge Eng2P8: Autonomous Driving • Cambridge EngMLMI7: Reinforcement Learning • Cambridge EngMLMI4: Advanced Machine Learning • Cambridge Eng3F8: Inference (Lab Leader) • Cambridge Eng4f13: Probabilistic Machine Learning • MISE Research Program (Teacher and Project Mentor) • University of Montreal IFT6135: Representation Learning (TA) 	2025 Easter (Spring) 2022, 2023, 2024 Lent (Winter) 2022, 2023, 2024 Lent (Winter) 2023 Lent (Winter) 2022, 2023 Michaelmas (Fall) 2021, 2022 Summer 2020 Spring 2019
----------	---	--

SUPERVISING	Post-docs:	
	• Ekdeep Singh Lubana, Harvard University (co-supervised with Hidenori Tanaka)	2024-2025
	• Fazl Barez, University of Oxford (co-supervised with Phil Torr)	2023-2024
	• Henry Ashton, University of Cambridge	2022-2023

PhD Students:		
• Jan Wehner, CISPA Helmholtz Center for Information Security (co-supervised with Mario Fritz)		2024-
• Alan Chan, Université de Montréal / Mila (co-supervised with Nicolas Le Roux)		2023-2025
• Bruno Mlodzeniec, University of Cambridge (co-supervised with Richard Turner)		2023-
• Shoaib Siddiqui, University of Cambridge		2022-
• Stephen Chung, University of Cambridge		2022-
• Usman Anwar, University of Cambridge (co-supervised with Jakob Foerster)		2022-
• Ethan Caballero, Université de Montréal / Mila (co-supervised with Irina Rish)		2022-
• Dmitrii Krasheninnikov, University of Cambridge		2021-
• Lauro Langosco, University of Cambridge		2021-
• Satyan Alex, University of Cambridge		2021-
• Nitarshan Rajkumar, University of Cambridge (co-supervised with Ferenc Huszár)		2021-
• Aryeh Englander, University of Maryland Baltimore County (co-supervised with I-Jeng Wang)		2021-

Master's Students:		
• Jose Miguel Lara Rangel, University of Cambridge MLMI (co-supervised with Usman Anwar)		2024
• Zsigmond Telek, University of Cambridge MLMI (co-supervised with Shoaib Siddiqui and Neel Alex)		2024
• Neela Aramandla, University of Cambridge MLMI (co-supervised with Neel Alex and Shoaib Siddiqui)		2024
• Tom Bush, University of Cambridge MLMI (co-supervised with Stephen Chung and Usman Anwar)		2024
• Zezhong Qin, University of Cambridge MEng (co-supervised with Bruno Mlodzeniec)		2023-2024
• Ognjen Stefanovic, University of Cambridge MEng (co-supervised with Ekdeep Singh Lubana and Alan Chan)		2023-2024
• Yawen Duan, University of Cambridge MLMI (co-supervised with Usman Anwar)		2023
• Thomas Coste, University of Cambridge MLMI (co-supervised with Usman Anwar)		2023
• William Baker, University of Cambridge MLMI (co-supervised with Lauro Langosco and Herbie Bradley)		2023
• Emilia Dordevic, University of Cambridge MLMI (co-supervised with Lauro Langosco)		2023
• Miguel Neves, University of Cambridge MLMI (co-supervised with Dmitrii Krasheninnikov and Ekdeep Singh Lubana)		2023

- Rudolf Laine, University of Cambridge CS (co-supervised with Lauro Langosco and Ferenc Huszár) 2022-2023
- Jason Brown, University of Cambridge MEng (co-supervised with Usman Anwar) 2022-2023
- Cindy Wu, University of Cambridge MEng (co-supervised with Ekdeep Singh Lubana and Robert Kirk) 2022-2023
- Alan Clark, University of Cambridge MLMI (co-supervised with Shoaib Siddiqui and Robert Kirk) 2022
- Andrei Alexandru, University of Cambridge (co-supervised with Lauro Langosco and Ferenc Huszár) 2021-2022
- Ethan Caballero, University of Montreal (co-supervised with Irina Rish) 2021-2022
- Yulong Lin, University of Cambridge MEng (co-supervised with Dmitrii Krasheninnikov and Robert Mullins) 2021-2022

Research Assistants and Interns:

- Vinaya Sivakumar 2026
- Justin Miller 2026
- Gideon Futerman (co-supervised with David Duvenaud, Jan Kulveit and Raymond Douglas) 2025
- Severin Field (co-supervised with David Duvenaud, Jan Kulveit and Raymond Douglas) 2025
- Max Ramsahoya (co-supervised with David Duvenaud, Jan Kulveit and Raymond Douglas) 2025
- Daniel Reuter 2025
- Lily Stelling 2025
- Brandon Jaipersaud (co-supervised with Ekdeep Lubana) 2025
- Yulu Pi (co-supervised with Alan Chan) 2024
- Michael Lan (co-supervised with Fazl Barez) 2024
- Minseon Kim (co-supervised with Fazl Barez) 2024
- Luke Marks (co-supervised with Fazl Barez) 2024
- Clement Neo (co-supervised with Fazl Barez) 2024
- Abhinav Menon (co-supervised with Ekdeep Lubana) 2024
- Joschka Braun (co-supervised with Dmitrii Krasheninnikov) 2024
- Carson Ezell (co-supervised with Alan Chan) 2024
- Kaivu Hariharan (co-supervised with Shoaib Siddiqui) 2024
- Itamar Pres (co-supervised with Ekdeep Lubana) 2024
- Joe Kwon (co-supervised with Usman Anwar and Dmitrii Krasheninnikov) 2024
- Karim Abdel Sadek (co-supervised with Usman Anwar) 2024
- Arturo Villacañas (co-supervised with Ekdeep Lubana and Usman Anwar) 2024
- Alexandra Bates (co-supervised with Lauro Langosco) 2023
- Ben Bucknall (co-supervised with Alan Chan) 2023
- Lexin Zhou (co-supervised with Gabriel Recchia) 2023
- Gabe Mukobi (co-supervised with Alan Chan) 2023
- Matthew Farrugia-Roberts (co-supervised with Usman Anwar) 2023
- Samyak Jain (co-supervised with Ekdeep Lubana) 2023
- Diego Dorn (co-supervised with Neel Alex) 2023
- Jesse Hoogland (co-supervised with Lauro Langosco) 2023
- Robert Klassert (co-supervised with Usman Anwar) 2023
- Ivan Anokhin (co-supervised with Stephen Chung) 2023
- Usman Anwar 2022
- Yawen Duan (co-supervised with Adam Gleave) 2021-2022
- Alexander Davies (co-supervised with Lauro Langosco) 2022

- Egor Krasheninnikov 2022-2024

Visitors:

- Jakub Vrabel, Brno University of Technology 2024
- Stephan Rabanser, University of Toronto 2023
- Rachel Freedman, University of California, Berkeley 2023
- Micah Carroll, University of California, Berkeley 2023
- Alan Chan, Université de Montréal / Mila 2022
- Sina Däubener, Ruhr-Universität Bochum 2022
- Joar Skalse, University of Oxford 2022
- Niki Howe, Université de Montréal / Mila 2022

SERVICE	<ul style="list-style-type: none"> • Organizer: NeurIPS Workshop On Socially Responsible LLM Research (SoLaR) 2023, 2024 • Organizer: San Francisco Alignment Workshop 2023 • Organizer: NeurIPS AI Safety Social 2022 • Organizer: ICML Workshop on Invertible Neural Nets and Normalizing Flows (INNF) 2019, 2020, 2021 • Organizer: AI Safety Unconference at NeurIPS 2018, 2019, 2022 • Organizer: NeurIPS Effective Altruism Social 2019 • AI Safety Index grading panel 2024, 2025 • Mila Student Applications Review Committee 2019 • Research Advisor, AI Safety Camp 2018-2020 • Organizer: NeurIPS workshop on Aligned Artificial Intelligence 2017 • Mila Lab Representative 2016 • Organizer: Mila reading groups on AI ethics, AI safety, Human Compatible AI, Radical Markets 2016-2020
MEDIA	<ul style="list-style-type: none"> • Initiator: Statement on AI Risk. The full contents of this statement are: “<i>Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.</i>” It has been signed by over 100 AI Professors, including Geoffrey Hinton and Yoshua Bengio. • TV Panelist: ITV’s Good Morning Britain “Could AI lead to the extinction of humanity?” • Author: New Scientist comment “Why do some AI researchers dismiss the potential risks to humanity?” • Interviewee: Nature “What counts as plagiarism? AI-generated papers pose new risks” • Op Ed: USA Today “No, AI isn’t inevitable. We should stop it while we can.” • Interviewee: Politico “How to pitch voters on the AI apocalypse” • Interviewee: MIT Technology Review “How AGI became the most consequential conspiracy theory of our time” • Podcast Guest: METAMIND AI “We’re NOT On Track to Solve Alignment: Here’s What Happens Next” • Podcast Guest: Le Pod with Leighton Woodhouse and Lee Fang • Podcast Guest: Robert Wright’s Nonzero Two Visions of AI Apocalypse • Interviewee: Daily Mail “Plunged into darkness while the oceans boil: How Mark Zuckerberg’s master plan will ‘lead to end of humanity’” • Interviewee: Social Studies “The Slow Way” • Interviewee: Business Insider “AI hype is crashing into reality. Stay calm.” • Interviewee: Business Insider “AI could lead to human extinction, says report commissioned by US State Department” • Interviewee: Le Journal de Québec “Ça pourrait mener à la mort de tous les humains: des experts en IA expliquent pourquoi la science-fiction est à nos portes”

- Interviewee: LeadDev Are you paying the AI competence penalty?
- Interviewee: MIT Technology Review “Forcing LLMs to be evil during training can make them nicer in the long run”
- Interviewee: The Wall Street Journal “How Worried Should We Be About AI’s Threat to Humanity? Even Tech Leaders Can’t Agree.”
- Interviewee: FastCompany “Military AI is here. Some experts are worried.”
- Interviewee: The Guardian US “‘Embrace it or risk obsolescence’: how will AI jobs affect Hollywood?”
- Interviewee: New Scientist “AI chatbots become more sycophantic as they get more advanced”
- Podcast Guest: Earthlings Podcast *The Fast Future of AI with David Krueger*
- Interviewee: Vox Future Perfect “The \$1 billion gamble to ensure AI doesn’t destroy humanity”
- TV Panelist: Al Jazeera’s Inside Story “Can regulating artificial intelligence suppress innovation?”
- Interviewee: Epsilooon “IA : Et maintenant, elle nous ment”
- Interviewee: Epsilooon “ChatGPT : ce n’est que le début”
- TV Interviewee: WCCO’s Good Question (CBS Minnesota) “How concerned should we be about extinction from AI?”
- Interviewee: Tencent News Periscope “AI horror! How long until human extinction? Initiator of the “AI Risk Statement”: Maybe it will be decades”
- Interviewee: France 24 article “Comment un monstre de Lovecraft est devenu un symbole du côté obscur des IA comme ChatGPT”
- TV Panelist: Al Jazeera’s Inside Story “Does Artificial Intelligence pose the risk of human extinction?”
- Interviewee: Associated Press article: “Artificial intelligence raises risk of extinction, experts say in new warning”
- Author: ai@cam “How can AI safety research reduce the risks of AI?”
- Interviewee: France 24 article “ChatGPT: mettre l’IA sur pause, ‘un enjeu existentiel’?”
- Podcast Guest: The Inside View “AI Alignment”
- Interviewee: PC Pro “Will AI kill us all? Serious minds think it might”
- Podcast Guest: Towards Data Science “Managing the incentives of AI”
- Podcast Guest: Future of Life “2018 AI Breakthroughs and Challenges”
- Volunteer Contributor: Graphite Publications: Series on AI
- Volunteer Reporter: KBOO Community Radio

INVITED TALKS / PANELS, ETC.	• University of California, Berkeley (Kavli speaker series)	2025
	• Governing Explosive Growth seminar	2025
	• MATS career fair	2025
	• PauseCon	2025
	• StopAI	2025
	• University of California, Berkeley (CHAI)	2025
	• NPTEL-NOC IITM	2025
	• Johns Hopkins	2025
	• Towards Safe and Trustworthy AI Agents Workshop at NeurIPS	2024
	• AI Safety Student Team	2024
	• Northeastern - David Bau’s group	2024
	• Harvard - Hidenori Tanaka’s group	2024
	• Harvard - Himabindu Lakkaraju’s group	2024
	• Harvard - Martin Wattenberg’s group	2024
	• Princeton Safety & Alignment Seminar	2024
	• New England Mechanistic Interpretability (NEMI) Workshop	2024

- ICM workshop on Models of Human Feedback for AI Alignment 2024
- ICM AI Safety Social: Navigating Misuse, Ethical Challenges, and Systemic Risks on Models of Human 2024
- Westminster eForum policy conference: Priorities for AI policy and regulation in the UK 2024
- Vienna Alignment Workshop 2024
- ACM India Summer School on Responsible and Safe AI 2024
- Human Aligned AI Summer School (Prague) 2024
- ERA Fellowship (Cambridge) 2024
- University of Oxford - FLAIR 2024
- University of Oxford - Torr Vision Group 2024
- UK Alignment Meetup 2024
- GovAI Winter Fellowship program 2024
- I Can't Believe It's Not Better Workshop at NeurIPS 2023
- AI Alignment Workshop 2023
- Future of Life Institute Existential Safety Community Member Meeting 2023
- Cohere UK Forum on Addressing Deployment Risks of Generative AI Systems 2023
- MSR New England ML Series 2023
- The Safe and Trustworthy AI Workshop 2023
- AI Safety Hub 2023
- Entrepreneur First 2023
- World AI Conference 2023
- Trustworthy and Responsible AI Conference 2023
- Stanford Existential Risks Initiative ML Alignment Theory Scholars (SERI-Mats) 2023
- Safe and Trusted AI Summer School 2023
- CHAI workshop 2023
- BAAI Alignment Forum 2023
- Tsinghua University 2023
- The Institute for AI Industry Research (AIR) 2023
- Effective Altruism Global (EAG): London 2023
- Makerere University 2023
- AGI Safety Fundamentals (AGISF) 2023
- Effective Altruism Global x (EAGx): Cambridge 2023
- Symposium on AGI Safety 2023
- University of Amsterdam (AMLAB seminar) 2023
- Center for AI Safety Philosophy Fellowship 2023
- NeurIPS ML Safety Workshop 2022
- Berkeley AI Research (BAIR) 2022
- 4th Scaling Laws Workshop 2022
- Cambridge AI Safety Hub 2022
- University of Toronto 2022
- University of Utah 2022
- University of Edinburgh 2022
- Stanford Existential Risks Initiative ML Alignment Theory Scholars (SERI-Mats) 2022
- NSF Convergence Accelerator Lightning Talk 2022
- 80,000 hours 2022
- Machine Learning Safety Scholars (MLSS) 2022
- Human-Aligned AI Summer School (HAAISS) 2022
- Johns Hopkins University Applied Physics Laboratory (APL) 2022
- Center for Security and Emerging Technology (CSET) 2022
- ELLIS Summer School 2022
- Concordia Chinese AI Safety Speaker Series 2022

	<ul style="list-style-type: none"> • Sea AI Lab • Symposium on AGI Safety • Oxford Artificial Intelligence Society • Cambridge Conference on Catastrophic Risk (CCCR, panelist) • 3rd Scaling Laws Workshop • 2nd Scaling Laws Workshop • Mila AGI Debates • DeepMind / Future of Humanity Institute AI safety Seminar • Vector Institute • AI Safety Support (AISS) Discussion Days • Center for Human-Compatible AI (CHAI) virtual workshop • HydroQuebec symposium 3i • Beneficial AGI (BAGI) conference • Montreal AI Ethics Institute (MAIEI) • EA Sherbrooke • Reed College • Ruhr University Bochum • Effective Altruism Global X (EAGx) Netherlands • Oxford Machine Learning groups • John Abbot College (a CEGEP) • CIFAR Deep Learning Summer School • Ottawa Machine Learning Meetup • Montreal Deep Learning Meetup • University of Montreal Undergraduate Student Association • IBM Research Presentation: RNN Regularization • Samsung workshop 	2022 2022 2022 2022 2022 2021 2021 2020 2020 2020 2019 2019 2018 2018 2018 2018 2018 2018 2018 2017 2017 2017 2016, 2017 2015 2015 2015
REVIEWING	<ul style="list-style-type: none"> • NeurIPS (top reviewer 2019, AC 2023-) • ICLR (top reviewer 2019, 2020) • ICLR (outstanding/notable reviewer award 2021, 2023) • AIES • FAccT • AAAI (meta-reviewer) • RLC (senior reviewer) • SaTML (notable reviewer award) • Patterns • MIT and Taylor Francis • CAIS Compute Cluster • CVPR • Artificial Intelligence Journal • JMLR • Montreal AI Symposium (MAIS) • Vitalik Buterin PhD Fellowship in AI Existential Safety • Machine Learning Reproducibility Challenge • NeurIPS Workshop on Pre-registration in ML • NeurIPS Workshop on Distribution Shifts • NeurIPS Workshop on Tackling Climate Change with Machine Learning • NeurIPS Workshop on Bayesian Deep Learning 	2019, 2020, 2022- 2019 - 2024 2021 - 2023, 2025 2023, 2024 2024 2023, 2024 2024 2023, 2024 2023 2023 2023 2023 2023 2022, 2023 2017 2020, 2021 2021, 2022 2019, 2021 2021 2021 2019 2017-2019
ASSESSMENT	PhD committee member: <ul style="list-style-type: none"> • Francis Rhys Ward, Imperial College London 	2024

- Mantas Mazeika, University of Illinois Urbana-Champaign 2024
- Ekdeep Singh Lubana, University of Michigan 2024
- Tomek Korbak, University of Sussex 2023
- Eric Daxberger, University of Cambridge 2023
- Ushnish Sengupta, University of Cambridge 2023
- Antonia Marcu, University of Southampton 2023
- Andrew Foong, University of Cambridge 2022

Master's projects:

- Faye Zhao, University of Cambridge MEng 2024
- Harry Le Xuan, University of Cambridge MEng 2024
- Byron Su, University of Cambridge MEng 2024
- Muqing Xue, University of Cambridge MLMI 2023
- Luke Peart, University of Cambridge MEng 2023
- Tom Ryan, University of Cambridge MEng 2023
- Vivek Palaniappan, University of Cambridge MEng 2023
- Matthew Barker, University of Cambridge MEng 2023
- Haitz Sáez De Ocáriz Borde, University of Cambridge MLMI 2022
- Adrian Black, University of Cambridge MLMI 2022
- Patrik Gergely, University of Cambridge MLMI 2022
- Zinzan Gurney, University of Cambridge MEng 2022
- Rakshit Jha, University of Cambridge MEng 2022
- Lihao Jiao, University of Cambridge MEng 2022

REFEREES**Yoshua Bengio**

Professor of Computer Science and Scientific Director of Mila
University of Montreal
julie.mongeau@mila.quebec

Stuart Russell

Professor of Computer Science and Director of Kavli Center for Ethics, Science, and the Public
University of California, Berkeley
russell@berkeley.edu

Carl Rasmussen

Professor of Machine Learning and Head of the Computational and Biological Learning Lab (CBL)
University of Cambridge
cer54@cam.ac.uk

Aaron Courville

Associate Professor of Computer Science
University of Montreal
aaron.courville@gmail.com

David Duvenaud

Associate Professor of Computer Science and Statistics
University of Toronto
duvenaud@vectorinstitute.ai

Jan Leike

Alignment Team Lead
OpenAI
jan@leike.name

Joel Dogoe

Founder/Director
MISE Educational Program in Ghana
mise.foundation@gmail.com

Roland Memisevic

Senior Director
Qualcomm
roland.memisevic@gmail.com