

# Silviu Pitis

🇨🇦 Canadian citizen

✉ silviu.pitis@gmail.com

📷 spitis

🐦 @silviupitis

📄 Google Scholar

🏠 silviupitis.com

## Research Summary

---

My research interest lies in the normative design of general-purpose decision making agents: how should we design generalist agents that benefit society? What objectives should our agents pursue? My current work focuses on designing objectives and evaluations for language-enabled agents that serve multiple principals, and applies tools from language modeling, decision theory, social choice, neural networks and representation learning. My full research statement is available at <https://silviupitis.com/#research>.

## Education

---

### University of Toronto

PH.D. IN COMPUTER SCIENCE, 4.0/4.0. ADVISOR: JIMMY BA.

Toronto, ON

Sep 2018 - Dec 2024

### Georgia Institute of Technology

M.S. IN COMPUTER SCIENCE, 4.0/4.0

Atlanta, GA

Sep 2016 - Dec 2017

### Harvard Law School

J.D., *Magna Cum Laude*, 4.2/4.0. JOHN M. OLIN FELLOW IN LAW AND ECONOMICS.

Cambridge, MA

Sep 2011 - May 2014

### Schulich School of Business, York University

B.B.A., *With Distinction*, 8.4/9.0

Toronto, ON

Sep 2006 - May 2010

## Publications

---

*All papers accessible via my Google Scholar profile*

### CONFERENCE

- **Silviu Pitis**, Ziang Xiao, Nicolas Le Roux, Alessandro Sordoni. [Improving Context-Aware Preference Modeling for Language Models](#). Neural Information Processing Systems (NeurIPS). 2024.
- Yangjun Ruan, Honghua Dong, Andrew Wang, **Silviu Pitis**, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, Tatsunori Hashimoto. [Identifying the Risks of LM Agents with an LM-Emulated Sandbox](#). International Conference on Learning Representations (ICLR). 2024.
- **Silviu Pitis**. [Consistent Aggregation of Objectives with Diverse Time Preferences Requires Non-Markovian Rewards](#). Neural Information Processing Systems (NeurIPS). 2023.
- Andrei Muresanu, Yongchao Zhou, Ziwen Han, Keiran Paster, **Silviu Pitis**, Harris Chan, Jimmy Ba. [Large Language Models are Human-Level Prompt Engineers](#). International Conference on Learning Representations (ICLR). 2023.
- **Silviu Pitis**, Elliot Creager, Ajay Mandelkar, Animesh Garg. [MoCoDA: Model-based Counterfactual Data Augmentation](#). Neural Information Processing Systems (NeurIPS). 2022.
- **Silviu Pitis**, Elliot Creager, Animesh Garg. [Counterfactual Data Augmentation using Locally Factored Dynamics](#). Neural Information Processing Systems (NeurIPS). 2020. **Outstanding Paper** at the Object-Oriented Learning Workshop at ICML 2020, where it was featured as a contributed talk.
- **Silviu Pitis\***, Harris Chan\*, Stephen Zhao, Bradly Stadie, Jimmy Ba. [Maximum Entropy Gain Exploration for Long Horizon Multi-goal Reinforcement Learning](#). International Conference on Machine Learning (ICML). 2020. **Best Paper** at the Adaptive and Learning Agents Workshop at AAMAS 2020, where it was featured as a contributed talk.
- **Silviu Pitis\***, Harris Chan\*, Jimmy Ba. [An Inductive Bias for Distances: Neural Nets that Respect the Triangle Inequality](#). International Conference on Learning Representations (ICLR). 2020.
- **Silviu Pitis**, Michael Zhang. [Objective Social Choice: Using Auxiliary Information to Improve Voting Outcomes](#). International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). 2020.
- Kristopher De Asis, Alan Chan, **Silviu Pitis**, Daniel Graves, Richard Sutton. [Fixed-Horizon Temporal Difference Methods for Stable Reinforcement Learning](#). AAAI Conference on Artificial Intelligence (AAAI). 2020.
- **Silviu Pitis**. [Rethinking the Discount Factor in Reinforcement Learning: A Decision Theoretic Approach](#). AAAI Conference on Artificial Intelligence (AAAI). 2019.
- **Silviu Pitis**. [Source Traces for Temporal Difference Learning](#). AAAI Conference on Artificial Intelligence (AAAI). 2018.
- **Silviu Pitis**. [Methods for Retrieving Alternative Contract Language Using a Prototype](#). International Conference on Law and Artificial Intelligence (ICAIL). 2017. (Oral, **Best Student Paper**).

## WORKSHOP

- Blair Yang, Fuyang Cui, Keiran Paster, Jimmy Ba, Pashootan Vaezipoor, **Silviu Pitis\***, Michael R. Zhang\*. [Report Cards: Qualitative Evaluation of Language Models Using Natural Language Summaries](#). Socially Responsible Language Modelling Research at NeurIPS. 2024.
- **Silviu Pitis**, Ziang Xiao, Alessandro Sordani. [Canonical Design for Language Agents using Natural Language Reward Models](#). AI meets Moral Philosophy and Moral Psychology Workshop at NeurIPS. 2023.
- **Silviu Pitis**. [Failure Modes of Learning Reward Models for LLMs and other Sequence Models](#). The Many Facets of Preference-based Learning at ICML. 2023.
- Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, **Silviu Pitis**, Roger Grosse, Jimmy Ba. [Calibrating Language Models via Augmented Prompt Ensembles](#). Deployment Challenges for Generative AI Workshop at ICML. 2023.
- Keiran Paster\*, **Silviu Pitis\***, Sheila A. McIlraith, Jimmy Ba. [Return Augmentation Gives Supervised RL Temporal Compositionality](#). Deep Reinforcement Learning Workshop at NeurIPS. Foundation Models for Decision Making Workshop at NeurIPS. 2022.
- Eric Xu, Jimmy Ba, **Silviu Pitis\***, Harris Chan\*. [Temporary Goals for Exploration](#). Deep Reinforcement Learning Workshop at NeurIPS. 2022.
- **Silviu Pitis\***, Harris Chan\*, Jimmy Ba. [ProtoGE: Prototype Goal Encodings for Multi-goal Reinforcement Learning](#). The 4th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM). 2019.
- **Silviu Pitis**. [Reasoning for Reinforcement Learning](#). Hierarchical RL Workshop, NeurIPS. 2017.

## PREPRINTS / OLDER UNPUBLISHED WORK

- Silviu Pitis, Michael R. Zhang, Andrew Wang, Jimmy Ba. [Boosted Prompt Ensembles for Large Language Models](#). Arxiv 2023.
- Silviu Pitis. [An Alternative Arithmetic for Word Vector Analogies](#). 2016.
- Silviu Pitis. [Punitive Damages in International Trade](#). Advised by Mark Wu. 2014.
- Silviu Pitis. [Designing Optimal Takeover Defenses](#). Advised by Holger Spamann. 2013.
- Silviu Pitis. [Examining Expected Utility Theory from Descriptive & Prescriptive Perspectives](#). Advised by Archishman Chakraborty. 2010.

## OPEN SOURCE PROJECTS

- Silviu Pitis, Harris Chan, Stephen Zhao. [mrl: modular RL](#). [github.com/spitis/mrl](https://github.com/spitis/mrl). Github. 2020.
- Silviu Pitis. [Swoosh: A leaner, faster backend for Whoosh](#). [github.com/spitis/PyIndex](https://github.com/spitis/PyIndex). Github. 2016.

## Fellowships & Awards

---

- 2024 - 2025 **Superalignment Fast Grant**, OpenAI (150K USD for research on aggregating subjective preferences)
- 2023 - 2024 **SRI Graduate Fellowship**, Schwartz Reisman Institute for Technology and Society
- 2020 - 2023 **Canada Graduate Scholarship - Doctoral (CGS D)**, Natural Sciences and Engineering Research Council of Canada
- 2019 - 2024 **Vector Institute Research Grant**, Vector Institute
- 2019 - 2020 **Ontario Graduate Scholarship (OGS)**, Government of Ontario
- 2018 - 2022 **Faculty of Arts and Science Top (FAST) Doctoral Fellowship**, University of Toronto
- 2018 **Faculty of Arts and Science Entrance Scholarship**, University of Toronto
- 2017 **Don Berman Best Student Paper**, International Conference on Law and Artificial Intelligence
- 2013 - 2014 **John M. Olin Fellowship in Law & Economics**, Olin Center for Law & Economics (1 of 3 in HLS Class of 2014)
- 2011 - 2014 **Dean's Scholar Prize (8x)**, Harvard Law School (awarded to top 1-2 students in large classes)
- 2006 - 2010 **York University Renewable Entrance Scholarship**, York University
- 2006 **Invitee**, Canadian Mathematics Olympiad (1 of 60 in Canada)

## Teaching & Supervision

---

### INSTRUCTOR

- Introduction to Machine Learning (CSC311). Fall 2020. (with Roger Grosse, Chris Maddison, Juhan Bae).

### TEACHING ASSISTANT

- Neural Networks and Deep Learning (CSC413/2516). Spring 2020.
- Machine Learning and Data Mining (CSC411). Spring 2019.
- Introduction to Artificial Intelligence (CSC384). Fall 2018.

### RESEARCH SUPERVISION / MENTORING

- Blair Yang and Fuyang (Scott) Cui. [Report Cards: Qualitative Evaluation of Language Models Using Natural Language Summaries](#). 2023. (with Michael R. Zhang, Pashootan Vaezipoor).
- Andrew Wang. [Variable Rate Discounting as Implicit Bayesian Reinforcement Learning](#). 2021-2022.
- Eric Xu. [Temporary Goals for Exploration](#). 2021-2022. (with Harris Chan).

## RESEARCH SUPERVISION / MENTORING, CONTINUED

- Ian Jiang. Architectures for Learning Local Causal Graphs. FY2021. (with Elliot Creager, Animesh Garg).
- Andrew Gritsevskiy. Leveraging Heuristic Search in Goal-Conditioned RL. 2020-2021. (with Harris Chan).
- Stephen Zhao. Layer-Wise Contrastive Unsupervised Representation Learning. Summer-Fall 2019. (with Jimmy Ba).
- Kiarash Jamali. Metric Nearness Experiments for Deep Norms. 2019-2020. (with Harris Chan).
- Bohan Zhang. Towards Unifying Deterministic and Stochastic Deep Reinforcement Learning Algorithms. Spring 2019. (with Jimmy Ba).

## Service

---

### REVIEWER / PROGRAM COMMITTEE

- ICLR 2025
- NeurIPS 2024
- NeurIPS 2023
- NeurIPS 2022
- ICML 2022
- AAI 2022
- NeurIPS 2021
- ICML 2021
- ICLR 2021
- AAI 2021
- NeurIPS 2020
- ICML 2020
- IJCAI 2020
- AAI 2020
- IJCAI 2019
- NeurIPS 2019
- ICML 2019

## Industry/Professional Experience

---

### Microsoft Research

RESEARCH INTERN (LARGE LANGUAGE MODELS; WITH ALESSANDRO SORDONI, ZIANG XIAO, NICOLAS LE ROUX)

*Montreal, QC (Remote)*

*May 2023 - August 2024*

### Kirkland & Ellis LLP

ASSOCIATE (CORPORATE LAW; PRIVATE EQUITY, PUBLIC M&A AND CROSS-BORDER TRANSACTIONS)  
SUMMER ASSOCIATE

*New York, NY*

*Oct 2014 - Mar 2016*

*May 2013 - Aug 2013*

### Harvard University

RESEARCH ASSISTANT (LAW & ECONOMICS; WITH PROFESSOR LOUIS KAPLOW)

*Cambridge, MA*

*Feb 2012 - May 2013*

### Pokerstars

PROFESSIONAL ONLINE POKER PLAYER

*Toronto, ON*

*Jan 2008 - Aug 2011*

## Other

---

### AFFILIATIONS

- Schwartz Reisman Institute. 2023 to present.
- Vector Institute for Artificial Intelligence. 2018 to present.
- AAI. 2017 to 2021.
- New York Bar. 2015 to present.
- Recurse Center. Spring 2016.
- John M. Olin Center for Law and Economics. 2013 to 2014.

### PROGRAMMING

- Poker hand evaluator and Nash equilibrium solver (Java).
- Real-time multi-user whiteboard application (Javascript).
- Contract aggregator and multi-way diff tool (C#, F#).
- Proprietary stock valuation model (Python, Tensorflow).
- Search engine (Python, 2x faster than best Python alternative).
- Deep reinforcement learning framework (Python, Pytorch).

### PRIVATE TUTORING / COACHING

- Poker. 2008 to 2011.
- Management statistics (undergraduate level). 2008.
- Math and physics (gifted high school level). 2006 to 2008.

### ENTREPRENEURIAL & RELATED

- Smart Contract Developer/Auditor (independent). 2021 to 2022.
- NextAI, Scientist-in-Residence. 2019 to 2020.
- DealWIP (legal tech startup). Advisor, 2019.
- iAgree LLC (alt. dispute resolution startup). Consultant. 2012.
- SCOPE. Cofounder & Director. 2006 to 2009. Not-for-profit that recruited over 60 student volunteers and ran goal-setting workshops for over 2000 Toronto high school students.