CHANDLER SMITH

+44 774534742 | smith.18.chandler@gmail.com | linkedin.com/in/chandlerdsmith | github.com/chansmi | chandlersmith.me/

EDUCATION

University of Oxford — October 2025 – Present
DPhil in Engineering Science (Advisor: Prof. Philip Torr and Dr. Christian Schroeder de Witt) — *Oxford, UK*

Northeastern University — January 2021 – December 2023
M.S. in Computer Science — *GPA: 3.93/4.0*

Colby College — August 2014 – May 2018
B.A. in Independent Major - Social Justice, Minor: Chemistry

PROFESSIONAL EXPERIENCE

Research Analyst — June 2024 – Present
Cooperative AI Foundation — *Oxford, UK*

- Directed UK AISI Evals program, delivering 3 critical evaluations in persuasion, jailbreaking, and oversight subversion for AI safety assessment
- Led the NeurIPS 2024 Concordia competition in collaboration with Google DeepMind, managing contest proposal, implementation strategy, technical report, and platform engineering for over 250 participants
- Reviewed 40+ grant proposals, evaluating technical content, topic relevance, and potential impact

Technical Architect Fellow, Applied Research — June 2023 – September 2025
In-Q-Tel — *Washington, DC*

- Led research on a multi-agent systems report by providing expertise on generative AI for national security
- Developed three market analysis architectures presenting: AI infrastructure, Generative AI, and AI Assurance
- Investigated 20+ early-stage companies performing due diligence and technical validation for investment consideration

Research Scholar — January 2024 – June 2024
Machine Learning Alignment Research Scholars Program — *Berkeley, CA*

- Designed and implemented a multi-agent benchmark technique on the punitiveness of foundation models
- Developed AI assessment for SOTA generative models, executing 200+ experiments in game-theoretic environments

Applications Engineer — May 2022 – May 2023
Dimagi - United States Health — *Washington, DC*

- Implemented three divisional performance Objectives and Key Results (OKRs), improving site performance monitoring for the CommCare web application platform
- Served as Solutions Architect for 9 divisional projects, managing technical requests, feasibility research, product specification generation, and implementation proposals

SELECT PUBLICATIONS

Evaluating Generalization Capabilities of LLM-Based Agents in Mixed-Motive Scenarios Using Concordia [1]
NeurIPS 2025 Datasets & Benchmarks Track

- Lead author evaluating agent generalization capabilities across diverse social dilemmas and strategic scenarios
- Developed comprehensive benchmark for assessing LLM-based agent behavior in mixed-motive environments

MALT: Learning to Reason with Multi-Agent LLM Systems [2]
Conference on Language Modeling (COLM) 2025

- Pioneered novel approach combining Direct Policy Optimization (DPO) and Supervised Fine-tuning (SFT) to enhance multi-model inference performance
- Co-led theoretical framework development and experimental design for demonstrating improved reasoning capabilities

Multi-agent Risks from Advanced AI [3]
arXiv - Cooperative AI Foundation

- Core team member contributing to the development of a multi-agent risk report, authoring 5 case studies that demonstrated AI-AI manipulation, multi-agent security, emergent capabilities, and destabilizing dynamics

BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Best Practices [4]
NeurIPS 2024 Datasets & Benchmarks (Spotlight)

- Core team member evaluating 15+ RL and LM benchmarks to establish comprehensive assessment criteria
- Developed framework for analyzing benchmark effectiveness, including authoring the initial project proposal

Escalation Risks from Language Models in Military and Diplomatic Decision-Making [5]
FAccT 2024

- Designed and conducted simulations modeling AI agent behavior in international relations scenarios
- Analyzed conflict escalation patterns in multi-agent defense strategy scenarios, identifying critical risk factors

The Concordia Contest: Advancing the Cooperative Intelligence of Language Model Agents [6]

NeurIPS 2024 Competition Track

- Led and designed a novel competition challenging participants to develop cooperative language model agents capable of navigating complex, text-mediated environments with competing interests and communication challenges
- Managed end-to-end competition logistics, technical infrastructure, and evaluation framework for 250+ participants

Evaluating Language Model Character Traits [7]

EMNLP 2024 (Findings)

- Conducted 100+ experiments examining behavioral patterns in language models (LMs) across various traits
- Developed formal framework for analyzing LM behavioral consistency without anthropomorphization

Conflict and Cooperation: An Exploration of Safety Concerns in Multi-Agent Systems [8]

Masters Thesis Project: Northeastern University

- Conducted comprehensive analysis of technical AI Safety, multi-agent risks, cooperative AI, and modern MARL techniques
- Successfully defended thesis examining technical foundations and failure modes of Multi-Agent AI systems

The State of AI in Maine [9]

Institute for Experiential AI at Northeastern University

- Led Energy and Natural Resource sector analysis based on 50+ industry interviews, identifying key ML capabilities
- Synthesized findings into comprehensive report for Institute for Experiential AI and Roux Institute stakeholders

Technical Skills

Languages and Tools: Python, C++, Docker, W&B, GCP, AWS Suite, Sagemaker, Slurm, Flux
Libraries and Frameworks: Pytorch, Tensorflow, Sklearn, HuggingFace, Inspect, CUDA, ROCm, TRL, JAX

Honors, Awards, and Professional Affiliation

- Cooperative AI PhD Scholar: Full funding for DPhil at University of Oxford (2025-2029)
- Research Affiliate for Technical Governance: Oxford Martin School AI Governance Institute
- Foresight Institute AI Safety Grant Recipient: Multi-agent Security, Steganography, and AI Control (2024)
- Open Philanthropy Grant Recipient: Multi-agent Research and Career Development (2024)
- Reviewer: NeurIPS 2024 Benchmarks and Datasets Track, Harms and Risks of AI in the Military 2024
- Cooperative AI Hackathon: First Place (2023)

References

[1] **Chandler Smith**, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit Trivedi, Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A. Duéñez-Guzmán, John P Agapiou, Jayd Matyas, Danny Karmon, Tim Baarslag, Dylan Hadfield-Menell, Natasha Jaques, Jose Hernandez-Orallo, Joel Z Leibo, et al. Evaluating generalization capabilities of LLM-based agents in mixed-motive scenarios using concordia. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[2] Sumeet Ramesh Motwani, **Chandler Smith**, Rocktim Jyoti Das, Markian Rybchuk, Ronald Clark, Philip Torr, Fabio Pizzati, Ivan Laptev, and Christian Schroeder de Witt. Malt: Learning to reason with multi-agent llm systems. In *Conference on Language Modeling (COLM) 2025*, August 2025.

[3] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, **Chandler Smith**, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai, 2025.

[4] Anka Reuel, Amelia Hardy, **Chandler Smith**, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[5] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, **Chandler Smith**, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. *FAccT 2024: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

[6] **Chandler Smith**, Rakshit Trivedi, Jesse Clifton, Lewis Hammond, Akbir Khan, Sasha Vezhnevets, John P Agapiou, Edgar A. Duéñez-Guzmán, Jayd Matyas, Danny Karmon, Marwa Abdulhai, Dylan Hadfield-Menell, Natasha Jaques, Joel Z Leibo, Oliver Slumbers, Tim Baarslag, and Minsuk Chang. The concordia contest: Advancing the cooperative intelligence of language agents. In *NeurIPS 2024 Competition Track*, 2024.

[7] Francis Rhys Ward, Zejia Yang, Alex Jackson, Randy Brown, **Chandler Smith**, Grace Colverd, Louis Thomson, Raymond Douglas, Patrik Bartak, and Andrew Rowan. Evaluating language model character traits, 2024.

[8] **Chandler Smith** and Bruce Maxwell. Conflict and cooperation: An exploration of safety concerns in multi-agent systems. *Northeastern University*, 2023.

[9] Anna Fiorentino, Olivia Saucier, Tyler Lynch, and **Chandler Smith**. The state of ai in maine. *The Institute for Experiential AI at Northeastern University*, 2023.