Sandy Tanwisuth

Curriculum Vitae

kst@berkeley.edu
sandytanwisuth.web.app
sandguine
Google Scholar

Research and Engineering Experiences

09/2025 – 12/2025	**Independent Researcher**, Funded by Cooperative AI Foundation Top 1% Early Career Research Grant Recipient
06/2025 – 08/2025	**Research Scholar** Machine Learning Alignment & Theory Scholars Senior Collaborator: Richard Ngo
04/2025 - 05/2025	**Research Consultant** Softmax
10/2024 – 05/2025	**Research Intern** Center for Human Compatible Artificial Intelligence, UC Berkeley Senior Collaborator: Niklas Lauffer
01/2023 – 01/2024	**Independent Researcher** In collaboration with University of Maryland MARL and Autonomous Decision Making Reading Group, Multi-agent Learning Seminars, and Noise Bridge AI
08/2019 – 12/2021	**Graduate Student Researcher** University of California Berkeley
09/2017 – 07/2019	**Post-baccalaureate Research Engineer** California Institute of Technology Advisor: John O'Doherty, Human Reward and Decision Making Laboratory
11/2015 – 05/2017	**Honors Thesis Student and Undergraduate Research Assistant** Arizona State University, Advisor: Samuel McClure, Decision Neuroscience Laboratory

Current Research Projects

In preparation for ICML 2026, **Strategic Representation Learning**	**Strategic Abstraction for Multi-Agent Coordination**, In multi-agent systems, different behaviors can lead to identical outcomes, raising the question of which distinctions are strategically relevant. Prior work on Strategic Equivalence Relations (SER) formalized exact equivalence via hard best responses but required full co-policy access. We introduce approximate Strategic Equivalence Classes (SECs), defined via soft best-response similarity, and learnable directly from interaction data with Strategic InfoNCE, a contrastive objective that embeds policies by how they deform the ego agent's incentive landscape. We prove learnability with finite-sample guarantees, show that $\Delta\mathrm{SEC}$ provides value-loss certificates, and highlight that short-horizon equivalence may hide long-term divergence. We also establish conditions preventing collapse of strategically distinct behaviors. Together, these results extend SER into a tractable, model-free framework, ensuring abstractions preserve incentive-relevant distinctions which is an essential property for safe coordination in diverse and human-centered environments.
Responsibilities	First author: ideations, conceptualizations, implementations, and writing (with advisor guidance about 0.5 - 1.5 hours per week for 6 months, then carried it to completion on my own.)
In preparation for ICML 2026, **Wellfounded Arbitration**	**Abstention-Aware Learning for Safe and Pluralistic Alignment** Ensuring safety in learning systems requires more than accuracy; agents must know when acting risks harm or when perspectives conflict. We study abstention-aware learning as a principled mechanism for risk reduction and pluralistic alignment. Extending margin-preserving abstractions from contextual bandits to reinforcement learning with trajectory-aware rules that capture uncertainty amplification through dynamics. Our approach combines the Strategic Equivalence Relations formalism with contrastive learning to learn soft–best-response embeddings and horizon-sensitive certificates that signal when action is epistemically justified. To safeguard pluralism, we add a coalitional arbitration layer that defers when internal experts disagree, preserving diverse strategies and human values.
Responsibilities	First author: ideations, conceptualizations, formalization, and writing (with advisor guidance about 0.5 hours per week for 2 months.)

Education

Graduate Courses	University of California Berkeley, Berkeley, CA
Department	**Cognitive Sciences and Electrical Engineering & Computer Sciences**
Relevant Coursework	**Deep Reinforcement Learning, Theory of Multi-armed Bandits and Reinforcement Learning, Multi-agent Systems and Population Games, Statistical Learning Theory, Methods in Computational Modeling for Cognitive Science**
	CGPA: 3.75/4.00, EECS GPA: 4.00/4.00
Bachelor of Science	Barrett, The Honors College, Arizona State University, Tempe, AZ
	Mathematics and Statistics, Symbolic Systems, and Cognitive Science
	CGPA: 3.94/4.00, Summa Cum Laude, Thesis Advisor: Samuel McClure

Management Experiences

06/2024 – 09/2024	**Research Manager** Machine Learning Alignment & Theory Scholars
02/2024 – 06/2024	**Research Operation Assitant** Center for Human Compatible Artificial Intelligence
09/2017 – 07/2019	**Laboratory Manager** California Institute of Technology

Teaching Experiences

Spring 2020 & 2021	UC Berkeley **Undergraduate Computational Cognitive Neuroscience**
Summer 2020	UC Berkeley **Undergraduate Statistical Methods**

Publications

2025 **S Tanwisuth**, D Leja ***Uncertainty-Aware Policy-Preserving Abstractions with Abstention for One-Shot Decisions*** NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists. 2025.

2023 K Iigaya, S Yi, I Wahle, **S Tanwisuth**, L Cross, J O'Doherty. ***Neural mechanisms underlying the hierarchical construction of perceived aesthetic value.*** Nature Communications. 2023.

2022 J Colas, N Dundon, R Gerraty, N Saragosa-Harris, K Szymula, **S Tanwisuth**, J Tyszka, C van Geen, H Ju, A Toga, J Gold, D Bassett, C Hartley, D Shohamy, S Grafton, J O'Doherty ***Reinforcement learning with associative or discriminative generalization across states and actions: fMRI at 3 T and 7 T.*** Human Brain Mapping. 2022.

2022 E Pool, R Gera, A Fransen, O Perez, A Cremer, M Aleksic, **S Tanwisuth**, S Quail, A Ceceli, D Manfredi, G Nave, E Tricomi, B Balleine, T Schonberg, L Schwabe, J O'Doherty ***Determining the effects of training duration on the behavioral expression of habitual control in humans: a multilaboratory investigation.*** Learning and Memory. 2022.

2021 K Iigaya, S Yi, I Wahle, **S Tanwisuth**, J O'Doherty. ***Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features.*** Nature Human Behavior. 2021.

2019 A Sitharanjan, **S Tanwisuth**. ***Exploring Historical Self Play for Autocurricula Generation.*** Deep Reinforcement Learning, Decision Making, and Control. 2019. UC Berkeley CS285 Final Project Repository.

Service

09/2025 Reviewer, **Aligning Reinforcement Learning Experimentalists and Theorists Workshop at NeurIPS**

09/2025 Reviewer, **Structured Probabilistic Inference & Generative Modeling: Probabilistic Inference in the Era of Large Foundation Models at NeurIPS**

09/2025 Reviewer, **Algorithmic Collective Action Workshop at NeurIPS**

07/2025 Reviewer, **Socially Responsible Language Modeling Research Workshop at Conference on Language Modeling**

04/2025-06/2025 Program Committee, **Coordination and Cooperation in Multi-Agent Reinforcement Learning Workshop at Reinforcement Learning Conference**

06/2024 Reviewer, **Trustworthy Multi-modal Foundation Models and AI Agents at International Conference on Machine Learning**

05/2024 Reviewer, **Coordination and Cooperation in Multi-Agent Reinforcement Learning Workshop at Reinforcement Learning Conference**

04/2024-06/2024 Program Committee, **Coordination and Cooperation in Multi-Agent Reinforcement Learning Workshop at Reinforcement Learning Conference**

02/2024-06/2024 Workshop Organizer, **8th Annual Center for Human-Compatible AI Workshop**

08/2020-Present Pro Bono Graduate Admission Consultant, **Project SHORT**

07/2020 Volunteer, **International Conference on Machine Learning**

06/2020-07/2020 Developer Volunteer, **Neuromatch Academy**

04/2020 Volunteer, **International Conference of Learning and Representation**

Honors, Awards, Scholarships, and Fellowships

2025 **Top 1%, Cooperative AI Foundation Early Career Grant Recipient**

2025 **Finalists, Astera Institute Fall Residency**

2024 **Cooperative AI Foundation Summer School Fellow**

2024-2025 **Center for Human-compatible AI Visiting Scholar Fellowship**

2023 **Cooperative AI Summer School Fellow** (unable to attend due to immigration constraint)

2023 **OpenAI ChatGPT Plugin Hackathon Finalists, MarvinGPT: a ChatGPT plugin that imbues the conversational AI with emotional intelligence**

2020 **Nominee for Microsoft Research: Ada Lovelace Fellowship**

2018 **CRCNS – Mining and Modeling Neuroscience Data Fellow, Redwood Center for Theoretical Neuroscience**

2016 – 2017 **Andre Levard Mackey Computational Study Scholarship**

2015 – 2016 **Jerry Witosky Memorial Scholarship**

2013 – 2017 **Dean's List** (Every Semester Throughout the Undergraduate Study)