Justin Svegliato

Research Scientist @ UC Berkeley

Artificial Intelligence — Planning — Reinforcement Learning — Large Language Models — AI Safety

☎ (631) 560-6259 ✉ jsvegliato@berkeley.edu 🏠 justinsvegliato.com

🎓 justinsvegliato 🐦 justinsvegliato 💼 justinsvegliato 💻 justinsvegliato Ⓜ justinsvegliato

Education

2022 **MS/PhD in Computer Science**, **University of Massachusetts Amherst**
PhD Candidate with Distinction – NSF Graduate Research Fellowship

2014 **BS in Computer Science and Philosophy with a Mathematics Minor**, **Marist College**
4.0 GPA – Valedictorian – NSF Technology Full Scholarship

Research Experience

University of California Berkeley — Berkeley, CA
Research Scientist in Artificial Intelligence — November 2021 – Present
- Led theoretical and empirical AI research in planning, reinforcement learning, large language models, and AI safety
- Applied novel AI methods to domains such as prompt engineering, pick-and-place robotics, and autonomous navigation
- Published 5+ papers at top conferences (AIJ/ICRA/IROS/ICAPS/AAMAS) with 5+ papers under review (ICLR/ICRA/JRC)
- Supervised ~10 undergraduate/graduate students on AI research projects that led to 10+ publications
- Spearheaded key initiatives, most notably the launch of a highly competitive internship program that attracted 300+ applicants
- Organized a 3-day CHAI workshop addressing critical AI issues like alignment and existential risk that attracted 200+ attendees
- Recognized as a Best Paper Award Finalist at a AAAI workshop and an Honorable Mention at a NeurIPS workshop

University of Massachusetts Amherst — Amherst, MA
Research Assistant in Artificial Intelligence — June 2016 – November 2021
- Built technical AI approaches that solve fundamental problems in planning, reinforcement learning, and AI safety
- Implemented AI systems in Python/JavaScript/C++ and built visualization tools in HTML/CSS by leveraging common libraries
- Applied novel AI methods to robotic systems such as autonomous vehicles, planetary rovers, and planet observation satellites
- Published 10+ papers at top conferences (AAAI/IJCAI/ICRA/IROS/AAMAS/AIES)
- Mentored 5+ undergraduate/graduate students on AI research projects that resulted in 10+ publications
- Organized a 1-day IROS workshop dedicated to building and evaluating ethical robotic systems that attracted 50+ attendees
- Recognized with 5+ honors including a Distinguished Paper Award at AAAI and an NSF Graduate Research Fellowship

Industry Experience

Nissan — Sunnyvale, CA
Research Intern in Autonomous Driving — May 2018 – May 2019
- Filed a patent as the primary inventor on detecting, identifying, and handling exceptions encountered by autonomous vehicles during operation by leveraging hierarchical belief space planning
- Implemented an exception recovery system on an operational autonomous vehicle prototype that effectively resolves a range of obstructions including road blocks, parked cars, garbage trucks, and pedestrians
- Formulated a mathematical framework that enables an autonomous vehicle to recover from exceptions by solving a novel hierarchy comprised of a POMDP for exception management and a set of MDPs each for a different exception
- Constructed a mathematical framework that allows an autonomous vehicle to dynamically adjust its level of autonomy based on feedback from a manual operator to ensure safe operation

Goldman Sachs — New York, NY
Software Engineer in Big Data Engineering — July 2015 – April 2016
- Designed, developed, and deployed a scalable web application that displays real-time/historical log data needed to troubleshoot and analyze application/service behavior without requiring direct server access

Lead Software Engineer in Network Engineering — May 2012 – July 2015
- Led the development of a network device abstraction service that uses a domain-specific language to abstract away the operating system details of network devices like routers, switches, firewalls, and load balancers
- Led the development of a network device discovery service that manages a configuration inventory for network devices like routers, switches, firewalls, and load balancers
- Developed a virtual IP provisioning service for a globally redundant route injection system to enhance network reliability
- Built an extensible network connection framework with support for common network protocols like SSH, Telnet, and HTTP

Academic Service

Teaching	COMPSCI 683: Artificial Intelligence (2018/2020), COMPSCI 190: Introduction to Game Design (2011)
Organizer	CHAI 2023, ERS 2021
Metareviewer	AAAI 2024, AAAI 2023
Reviewer	ICRA 2024, AAAI Demo 2024, NeurIPS Workshop 2023, ICRA 2023, UAI 2022, IJCAI 2022, NeurIPS 2021, UAI 2021, IROS 2021, IJCAI 2021, ICRA 2021, AAAI 2021, IROS 2020, ICRA 2020, AAAI 2020, IEEE 2019a, IEEE 2019b

Honors

2023	**Spotlight Talk:**	Awarded to the top 5% of conference papers at ICLR
2023	**Honorable Mention:**	Awarded to top workshop papers on LLMs at the NeurIPS Workshop on ITIF
2023	**Best Paper Award Finalist:**	Awarded to top workshop papers on RLHF at the AAAI Workshop on SafeAI
2021	**Distinguished Paper Award:**	Awarded to top conference papers at AAAI
2019	**Distinguished Teaching Award Finalist:**	Awarded to a few top teaching assistants
2018	**NSF Graduate Research Fellowship:**	Awarded to high-potential PhD students early in their career
2018	**NSF Grant on Robust Intelligence:**	Awarded $450,000 to conduct research on decision making under uncertainty
2018	**PhD Candidate with Distinction:**	Awarded to top PhD students for excellence on the PhD candidacy qualifier exam
2016	**Victor Lesser Graduate Scholarship in AI:**	Awarded to a top incoming PhD student in AI
2014	**Valedictorian:**	Awarded to the top graduating student in college
2014	**Excellence Award in Computer Science:**	Awarded to the top graduating student in computer science
2014	**Excellence Award in Philosophy:**	Awarded to the top graduating student in philosophy
2014	**Intern of the Year:**	Awarded to the top graduating student for excellence in industry
2014	**Summa Cum Laude:**	Awarded to graduating students with at least a 3.85 GPA
2013	**Presidential Scholar Student Speaker:**	Selected to give 6 talks to top accepted prospective students
2010	**Deans' Circle:**	Admitted to an honors organization with only the top 3% of students
2010	**Deans' List:**	Awarded to students with at least a 3.6 GPA
2010	**NSF Technology Full Scholarship:**	Included a full scholarship for all years of college
2010	**Fiovranti Memorial Scholarship for Athletics:**	Awarded for excellence in cross country and track and field

Selected Publications

ICML 2024	**Active Teacher Selection for Reward Learning** – **In Submission**
IJCAI 2024	**Learning to Plan with Tree Search via Deep RL** – **In Submission**
ICRA 2024	**Ethically Compliant Autonomous Systems under Partial Observability** – **In Submission**
JRC 2024	**Fairness and Sequential Decision Making: Limits, Lessons, and Opportunities** – **In Submission**
ICLR 2024	**Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game** – **Spotlight Talk**
AAMAS 2024	**Defining Deception in Decision Making**
IROS 2023	**Formal Composition of Robotic Systems as Contract Programs**
AIJ 2022	**Competence-Aware Systems**
IROS 2022	**Selecting the Partial State Abstractions of MDPs: A Metareasoning Approach with Deep RL**
Dissertation 2022	**Metareasoning for Planning and Execution in Autonomous Systems**
ICAPS 2022	**Tuning the Hyperparameters of Anytime Planning: A Metareasoning Approach with Deep RL**
ICRA 2022	**Metareasoning for Safe Decision Making in Autonomous Systems**
IROS 2021	**Agent-Aware State Estimation for Autonomous Vehicles**
IROS 2021	**Improving Competence via Iterative State Space Refinement**
AIES 2021	**Ethically Compliant Planning within Moral Communities**
SoCS 2021	**On the Benefits of Randomly Adjusting Anytime Weighted A***
ICRA 2021	**Solving Markov Decision Processes with Partial State Abstractions**
AAAI 2021	**Ethically Compliant Sequential Decision Making** – **Distinguished Paper Award**
US 2020	**Introspective Autonomous Vehicle Operational Management**
ICRA 2020	**A Model-Free Approach to Meta-Level Control of Anytime Algorithms**
AAMAS 2020	**Learning to Optimize Autonomy in Competence-Aware Systems**
ECAI 2020	**An Integrated Approach to Moral Autonomous Systems**
IROS 2019	**Belief Space Metareasoning for Exception Recovery**
IJCAI 2018	**Meta-Level Control of Anytime Algorithms with Online Performance Prediction**

Technical Skills

Languages (Advanced)	Python, Java, JavaScript, HTML, CSS, SQL, NoSQL (MongoDB), Bash, LaTeX
Languages (Proficient)	TypeScript, PHP, Visual Basic
Languages (Basic)	C++, MATLAB, Maple, R, Haskell, Lisp, Scheme, Racket, Perl, Fortran, Prolog, z/OS Assembly
Libraries (Python)	ROS, Django, Gym, NumPy, SciPy, Scikit-Learn, Keras, Theano, PyTorch, Matplotlib, Seaborn, Tkinter
Libraries (Java)	JUnit, Mockito, Jackson, Gson, Log4j, SLF4j, Apache Commons, Apache Tomcat, Gradle, Maven, Ant
Libraries (JavaScript)	Node, npm, React, Flux, Redux, jQuery, D3, Chart, Underscore, Lodash, Moment, Grunt, Gulp
Libraries (HTML/CSS)	Bootstrap, Tailwind, Material Design, Material Design Lite
Editors	Visual Studio Code, IntelliJ, NetBeans, Eclipse, PyCharm, CLion, Code::Blocks, Overleaf, Sublime Text
Utilities	WandB, Jupyter, JIRA, Confluence, Git, SVN, Vim, Nano, Terminator, iTerm, Aqua Data Studio
Graphic Design	Adobe Photoshop/ImageReady, Blender, Astah Professional, Paint.NET, GIMP, Draw.io
Licenses	Private Pilot License (Glider Rating), Sailing License, Amateur Extra Radio License (Call Sign: K3GR)
Certifications	NAUI Scuba Diver Certification, NAUI EANx Diver Certification, AHA First Aid/CPR/AED Certification
Foreign Languages	Spanish (Intermediate Reading, Intermediate Writing, Beginner Speaking, Beginner Listening)