

Eugene Bagdasarian

140 Governors Dr, Office E451
Amherst, MA 01003
 eugene@umass.edu
 0000-0002-7994-6469
 people.cs.umass.edu/~eugene

Research Interests

I study security and privacy attack vectors in deployed and emerging AI systems. My research informs the design of these systems to be trustworthy, safe, ethical, and resilient to attacks.

Experience

- 2024 – Present **Assistant Professor of Computer Science**, *Manning College of Information and Computer Sciences*, University of Massachusetts Amherst
- 2023 – Present **Senior Research Scientist**, *Google Research*, part-time
- 2014 – 2016 **Software Engineer**, *Cisco Systems*

Education

- Cornell Tech, Cornell University**, New York, NY, USA
- 2016–2023 PhD in Computer Science. Advised by Vitaly Shmatikov and Deborah Estrin
- 2016–2019 MSc in Computer Science
- Bauman Moscow State Technical University**, Moscow, Russia
- 2009–2016 Engineer's degree in Computer Science, *summa cum laude*
- 2009–2013 BS in Computer Science, *summa cum laude*

Awards and Honors

- 2024 USENIX Security Distinguished Paper Award
- 2023 Cornell Tech PhD Excellence Award
- 2021 Apple Scholars in AI/ML PhD Fellowship
- 2019 Digital Life Initiative Doctoral Fellowship
- 2017 Bloomberg Data For Good Exchange Award
- 2017 Computer Science Dept TA Excellence Award
- 2011,'12,'13 Potanin Foundation Scholarship
- 2011,'12 Bauman University Academic Excellence Fellowship

Funding

- 2025 **Schmidt Sciences AI Safety Grant**, \$500,000
- Multi-agent Safety. Main PI, Co-PI Shlomo Zilberstein

Keynotes

- Oct 2025 **ACM CCS'25 AI Security Workshop**, Keynote on *Privacy and Security for Future AI Agents*
- Aug 2025 **SOUUPS'25 Societal & User-Centered Privacy in AI (SUPA) Workshop**, Keynote on *Designing Privacy-conscious AI Agents*
- May 2025 **IEEE S&P'25 Secure Generative AI Agents Workshop**, Contributed long talk on *Contextual Defenses for Privacy-conscious Agents*
- Mar 2025 **AAAI'25 Deployable AI Workshop**, Keynote on *Dangers in Inference-heavy AI Pipelines: Embeddings and Reasonings*
- Mar 2025 **AAAI'25 Privacy-Preserving AI Workshop**, Keynote on *Contextual Integrity for Privacy-conscious Agents*
- Dec 2023 **NeurIPS'23 Trojan Detection Challenge**, Keynote on *multi-modal attacks in visual language models*

Selected Publications

Lillian Tsai and **Eugene Bagdasarian**. Contextual agent security: A policy for every purpose. In *HotOS*, 2025.

Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. AirGapAgent: Protecting privacy-conscious conversational agents. In *CCS*, 2024.

Tingwei Zhang, Rishi Jha, **Eugene Bagdasaryan**, and Vitaly Shmatikov. Adversarial illusions in multi-modal embeddings. In *USENIX Security*, 2024.  **Distinguished Paper Award**.

Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *S&P*, 2022.

Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *USENIX Security*, 2021.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*, 2020.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.

Internships

- 2021 – 2021
May Aug **Research Intern**, Apple, Cupertino, CA, USA
Conducted research on federated learning and language models.
- 2020 – 2020
May Aug **Research Intern**, Google Research, New York, NY, USA
Researched local differential privacy and secure aggregation for federated analytics.

2018 – 2018	May Aug	Applied Scientist Intern , <i>Amazon</i> , Seattle, WA, USA Worked on a novel multi-service recommendations engine for Alexa.
2013 – 2014	Aug July	Software Engineering Intern , <i>Cisco Systems</i> , Boston, MA, USA Developed front-end and back-end for the SocialMiner data analytics web application.

Organizer

- Dec 2023 **NeurIPS'23 Workshop**, *Backdoors in Deep Learning: The Good, the Bad, and the Ugly*
Co-organizer of workshop on backdoor attacks and defenses in deep learning
- Dec 2018 **RecSys'18 Tutorial**, *Modularizing Deep Neural Network-Inspired Recommendation Algorithms*
In collaboration with Longqi Yang and Hongyi Wen

Media Coverage

- Apr 2023 **The Economist**, “It doesn’t take much to make machine-learning algorithms go awry”
- Oct 2022 **Pluralistic: Cory Doctorow**, “Backdooring a summarizerbot to shape opinion”
- Oct 2022 **Schneier on Security**, “Adversarial ML Attack that Secretly Gives a Language Model a Point of View”
- Dec 2021 **VentureBeat**, “Propaganda-as-a-service may be on the horizon if large language models are abused”
- Aug 2021 **ZDNet**, “Cornell University researchers discover ‘code-poisoning’ attack”
- Jun 2020 **Cornell Chronicle**, “Platform empowers users to control their personal data”

Advising

PhD Students

- 2024–Present Abhinav Kumar
- 2025–Present Dzung Pham (co-advised w Amir Houmansadr)
- 2025–Present June Jeong (co-advised w Amir Houmansadr)

Teaching Experience

- Fall 2025 COMPSCI 690F: Trustworthy and Responsible AI
- Fall'25, Spring'25, Fall'25 COMPSCI 692: AI Security Seminar
- Spring 2025 COMPSCI 360: Introduction to Security
- Spring 2017 CS 5450: Networked and Distributed Systems, TA, **Excellence Award**

Professional Activity

Conference Reviewing

S&P'26, ICLR'25, CCS'25, CCS'24, ICLR'24, ICLR'22, ICML'22, NeurIPS'21

Journal Reviewing

TMLR'22, IEEE T-IFS'22

Workshop Reviewing

FL4NLP@ACL'22, AdvML@ICML'22, MAISP@MobiSys'21

Department Service

- 2024–Present Co-Lead of AI Safety Initiative at UMass
2024–Present Organizer of the CICS Security and Privacy Seminar Series
2018–2019 Co-lead of the PhD Student at Cornell Tech (PACT) organization

Broadening Participation

- 2024–Present Co-Organizer of Pioneer Leaders in AI and Robotics Initiative

Invited Talks

- Sep 2025 **Northeastern University, Security Seminar**
Building Trustworthy Future AI Agents
- Mar 2025 **Brave, Research Seminar**
Designing privacy-conscious Agents
- Oct 2024 **ServiceNow, Research Seminar**
Designing privacy-conscious Agents
- Apr 2023 **Michigan CS, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Apr 2023 **Columbia CS, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Apr 2023 **BU CDS, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Mar 2023 **UW Allen School CSE, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Mar 2023 **McGill, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Feb 2023 **CISPA, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Feb 2023 **UMass CS, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?
- Jan 2023 **UCLA CS, Research Seminar**
Untrustworthy Machine Learning: How to Balance Security, Accuracy, and Privacy?

- Sep 2022 **Brave Software**, *Research Seminar*
Sparse federated analytics: location heatmaps and language tokenizations.
- Jul 2022 **Google Research**, *Google Federated Talks*
Sparse federated analytics: location heatmaps and language tokenizations.
- Mar 2022 **University of Chicago**, *The SAND Lab Talks*
Spinning Language Models: Propaganda-As-A-Service and Countermeasures.
- Jan 2022 **University of Cagliari**, *Machine Learning Security Seminar Series*
Spinning Language Models: Propaganda-As-A-Service and Countermeasures.
- Jan 2022 **Samsung AI Center Cambridge**, *Invited Talk Series*
Evaluating privacy preserving techniques in machine learning.
- Dec 2021 **University College London**, *Privacy and Security in ML Interest Group*
Blind Backdoors in Deep Learning Models.
- Nov 2021 **University of Cambridge**, *Computer Laboratory Security Seminar*
Blind Backdoors in Deep Learning Models.
- Sep 2021 **Telefonica Research**, *Research Seminar*
Evaluating privacy preserving techniques in machine learning.
- Jan 2021 **Microsoft**, *Applied Research Invited Talk Series*
Evaluating privacy preserving techniques in machine learning.
- Jun 2020 **Google Research**, *Google Federated Talks*
Salvaging federated learning with local adaptation.
- Feb 2020 **Cornell Tech**, *Digital Live Initiative*
Evaluating privacy preserving techniques in machine learning.

All Publications

Conference Publications

Ren Yi, Octavian Suciu, Adria Gascon, Sarah Meiklejohn, **Eugene Bagdasarian**, and Marco Gruteser. Privacy reasoning in ambiguous contexts. In *NeurIPS*, 2025.

Tingwei Zhang, Collin Zhang, John X. Morris, **Eugene Bagdasarian**, and Vitaly Shmatikov. Self-interpreting adversarial images. In *USENIX Security*, 2025.

Ali Naseh, Jaechul Roh, **Eugene Bagdasarian**, and Amir Houmansadr. Backdooring bias into text-to-image models. In *USENIX Security*, 2025.

Eugene Bagdasarian and Vitaly Shmatikov. Mithridates: Auditing and boosting backdoor resistance of machine learning pipelines. In *CCS*, 2024.

Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. AirGapAgent: Protecting privacy-conscious conversational agents. In *CCS*, 2024.

Tingwei Zhang, Rishi Jha, **Eugene Bagdasaryan**, and Vitaly Shmatikov. Adversarial illusions in multi-modal embeddings. In *USENIX Security*, 2024.
🏆 Distinguished Paper Award.

Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *S&P*, 2022.

Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *USENIX Security*, 2021.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*, 2020.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.

Zhiming Shen, Zhen Sun, Gur-Eyal Sela, **Eugene Bagdasaryan**, Christina Delimitrou, Robbert Van Renesse, and Hakim Weatherspoon. X-containers: Breaking down barriers to improve performance and isolation of cloud-native containers. In *ASPLOS*, 2019.

Longqi Yang, **Eugene Bagdasaryan**, Joshua Gruenstein, Cheng-Kang Hsieh, and Deborah Estrin. Openrec: A modular framework for extensible and adaptable recommendation algorithms. In *WSDM*, 2018.

Longqi Yang, **Eugene Bagdasaryan**, and Hongyi Wen. Modularizing deep neural network-inspired recommendation algorithms. In *RecSys*, 2018.

Journal Publications

Sahra Ghalebikesabi, **Eugene Bagdasarian**, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, Pushmeet Kohli, Po-Sen Huang, and Borja Balle. Privacy awareness for information-sharing assistants: A case-study on form-filling with contextual integrity. *TMLR*, 2025.

Eugene Bagdasaryan, Peter Kairouz, Stefan Mellem, Adrià Gascón, Kallista Bonawitz, Deborah Estrin, and Marco Gruteser. Towards sparse federated analytics: Location heatmaps under distributed differential privacy with secure aggregation. In *PETS*, 2022.

Workshop Papers

Lillian Tsai and **Eugene Bagdasarian**. Contextual agent security: A policy for every purpose. In *HotOS*, 2025.

Eugene Bagdasaryan, Congzheng Song, Rogier van Dalen, Matt Seigel, and Áine Cahill. Training a tokenizer for free with private federated learning. In *FL4NLP at ACL*, 2022.

Eugene Bagdasaryan, Griffin Berlstein, Jason Waterman, Eleanor Birrell, Nate Foster, Fred B Schneider, and Deborah Estrin. Ancile: Enhancing privacy for ubiquitous computing with use-based privacy. In *WPES at CCS*, 2019.

Preprints

Mason Nakamura, Abhinav Kumar, Saaduddin Mahmud, Sahar Abdelnabi, Shlomo Zilberstein, and **Eugene Bagdasarian**. Terrarium: Revisiting the blackboard for multi-agent safety, privacy, and security studies. *arXiv preprint arXiv:2510.14312*, 2025.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Amir Houmansadr, and **Eugene Bagdasarian**. Throttling web agents using reasoning gates. *arXiv preprint arXiv:2509.01619*, 2025.

Hyeyun Jeong, Mohammadreza Teymoorianfard, Abhinav Kumar, Amir Houmansadr, and **Eugene Bagdasarian**. Network-level prompt and trait leakage in local research agents. *arXiv preprint arXiv:2508.20282*, 2025.

Dzung Pham, Peter Kairouz, Niloofar Mireshghallah, **Eugene Bagdasarian**, Chau Minh Pham, and Amir Houmansadr. Can large language models really recognize your name? *arXiv preprint arXiv:2505.14549*, 2025.

Sahar Abdelnabi, Amr Gomaa, **Eugene Bagdasarian**, Per Ola Kristensson, and Reza Shokri. Firewalls to secure dynamic LLM agentic networks. *arXiv preprint arXiv:2502.01822*, 2025.

Kleomenis Katevas, **Eugene Bagdasaryan**, Jason Waterman, Mohamad Mounir Safadieh, Eleanor Birrell, Hamed Haddadi, and Deborah Estrin. Policy-based federated learning. *Preprint*, 2020.

Tao Yu, **Eugene Bagdasaryan**, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *Preprint*, 2020.

Jonathan Behrens, Ken Birman, Sagar Jha, Matthew Milano, Edward Tremel, **Eugene Bagdasaryan**, Theo Gkountouvas, Weijia Song, and Robbert Van Renesse. Derecho: Group communication at the speed of light. Technical report, Cornell University, 2016.