140 Governors Dr, CS 304
Amherst, MA, 01003
✉ *eugene@umass.edu*
🌐 *people.cs.umass.edu/˜eugene/*
iD *0000-0002-7994-6469*
Old spelling: Eugene Bagdasaryan

Eugene Bagdasarian

Research Interests

Eugene studies security and privacy in emerging machine learning technologies under real-life conditions and attacks. The main goal of his research is to build ML and GenAI-based systems that are ethical, safe, and private by design.

Experience

2024 – Present Aug	**Assistant Professor of Computer Science**, *Manning College of Information and Computer Sciences*, University of Massachusetts Amherst
2023 – Present Aug	**Research Scientist**, *Google Research*, Cambridge, MA
2014 – 2016 Sep Jul	**Software Engineer**, *Cisco Innovation Center*, Moscow, Russia

Education

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

Awards and Honors

2024	Distinguished Paper Award at USENIX Security
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

Internships

2021 May – 2021 Aug **Research Intern**, *Apple*, Cupertino, CA, USA
Conducted research on federated learning and language models.

2020 May – 2020 Aug **Research Intern**, *Google Research*, New York, NY, USA
Researched local differential privacy and secure aggregation for federated analytics.

2018 May – 2018 Aug **Applied Scientist Intern**, *Amazon*, Seattle, WA, USA
Worked on a novel multi-service recommendations engine for Alexa.

2013 Aug – 2014 July **Software Engineering Intern**, *Cisco Systems*, Boston, MA, USA
Developed front-end and back-end for the SocialMiner data analytics web application.

2012 Dec – 2013 Apr **Intern**, *Deloitte*, Moscow, Russia
Performed data analytics tasks for the audit department.

Publications

Conference Publications

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.

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

Tingwei Zhang, Rishi D 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

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 and Preprints

Zhao Cheng, Diane Wan, Matthew Abueg, Sahra Ghalebikesabi, Ren Yi, **Eugene Bagdasarian**, Borja Balle, Stefan Mellem, and Shawn O'Banion. CI-Bench: Benchmarking contextual integrity of AI assistants on synthetic data. *Preprint*, 2024.

Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and **Eugene Bagdasaryan**. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI. *Preprint*, 2024.

Tingwei Zhang, Collin Zhang, John X Morris, **Eugene Bagdasarian**, and Vitaly Shmatikov. Soft prompts go hard: Steering visual language models with hidden meta-instructions. *Preprint*, 2024.

Ali Naseh, Jaechul Roh, **Eugene Bagdasarian**, and Amir Houmansadr. Injecting bias in text-to-image models via composite-trigger backdoors. *arXiv preprint arXiv:2406.15213*, 2024.

Sahra Ghalebikesabi, **Eugene Bagdasaryan**, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, et al. Operationalizing contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373*, 2024.

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.

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.

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.

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.

Tutorials

Dec 2018 **RecSys 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”

Mentoring Experience

Undergraduate Students

2019–2020 Mohamad Mounir Safadieh, Vassar College → Apple

2017–2019 Griffin Berlstein, Vassar College → Cornell PhD program

Masters Students

2021-2022 Anastasia Sorokina, Cornell Tech

2020 Pargol Gheissari, Cornell Tech → Palantir

2020 Kuan-Ting Liu, Cornell Tech → Facebook

2020 Calvin Li, Cornell Tech → Evernorth

2020 Chinmay Bhat, Cornell Tech → Shift Technology

2020 Surya Omesh, Cornell Tech → Bloomberg

2020 Saloni Gandhi, Cornell Tech → Twitter

2020 Devansh Gosalia, Cornell Tech → OppFi

2019 Rony Krell, Cornell Tech → UnitedHealth Group

Teaching Experience

Fall 2024 COMPSCI 692PA: Advanced Topics on Security and Privacy for Generative Models

Spring 2022 CS 5436/INFO 5303: Privacy in the Digital Age, Teaching Assistant, Part-time

Spring 2021 CS 5436/INFO 5303: Privacy in the Digital Age, Teaching Assistant, Part-time

Spring 2020 CS 5450: Networked and Distributed Systems, Teaching Assistant, Part-time

Fall 2018 CS 5450: Networked and Distributed Systems, Teaching Assistant, Part-time

Spring 2017 CS 5450: Networked and Distributed Systems, Teaching Assistant, Full-time, **Excellence Award**

Fall 2016 CS 4320: Introduction to Database Systems, Teaching Assistant, Full-time

Professional Activity

Program Committee

CCS'25, CCS'24

Reviewer

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

2018–2019 Co-lead of the PhD Student at Cornell Tech (PACT) organization

Invited Talks

Dec 2023 **NeurIPS**, *TrojAI Competition Invited Talk*
Multi-modal Attacks on Visual Language Models

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.