Ankur Taly

Email: ankur.taly@gmail.com
Phone: (732) 513-5690
Homepage: http://theory.stanford.edu/∼ataly

Summary

Tech lead and researcher with experience spanning **machine learning**, **natural language processing**, **computer security**, and **formal methods**. My current work at Google is focussed on improving grounding and factuality of Large Language Models (LLMs). Prior work at Fiddler labs (2019 — 2020) and Google Brain (2016 — 2019) was focussed on interpreting and explaining machine learning models. Prior to this (2007—2016), I worked on computer security — identity and authorization, protocol verification, and programming language security. I have published papers in top-tier venues of each of these fields, cumulating over **9500 citations**.

Employment

- **Google Inc., Sunnyvale, CA, USA**

  Tech Lead Manager, Google Cloud (Sep. 2023 — present)

  LLM Grounding and Factuality

  – Lead a cross-team effort focussed on quality of LLM deployments in Google Cloud. Manage two broad workstreams – (1) post-hoc evaluation and guardrails to improve factuality, and (2) recipes (e.g., retrieval augmented generation, corroborate and revise, etc.) to enable factual generation.

  – Responsibilities include overseeing all human evaluations of LLM deployments in Cloud, citation generation for LLM responses, developing auto raters for quality evaluation, identifying quality gaps and using them to steer upstream modeling of LLMs and recommending new response generation recipes for downstream applications.

  – Recent launch from my immediate team: CheckGrounding API to determine whether a response is grounded in a set of facts, and generate citations.

  Staff Research Scientist, Google Cloud (Jul. 2020 — Sep. 2023)

  Training Data Curation

  – Developed techniques for quantifying training data influence in language models (Bert, RoBERTa), identifying mislabeled examples, and curating finetuning sets to minimize forgetting of pretraining knowledge (**refer Publications [7, 33]**).

  – Helped several Cloud AI teams with curating training data sets, and saving on human labeling costs. Recent launch: Selective labeling tool in Document AI.

  Interpreting Machine Learning models

  – Developed inherently interpretable models, and techniques for analyzing machine learning (ML) models. Launched as part of https://pair-code.github.io/lit/ (**refer Publications [2, 8, 9]**).

  – Consulted as an explainability expert across several customer engagements in Cloud AI.

- **Fiddler Labs, Mountain View, CA, USA** May 2019 — Jul. 2020

  Head of Data Science

  – Oversaw all data science efforts, and managed the data science team.

  – Drove research on pushing the state of the art on explainability methods. Conceived and developed Fiddler's Slice and Explain framework (**refer Publications [10, 11]**).

  – Served as a technical expert in all sales conversations, and helped set up company wide product strategy and roadmap.

  - Evangelized explainable AI technology across industry and academia — co-organized tutorials on explainable AI at several top-tier academic conferences, and presented Fiddler's explainable AI technology at 10+ companies (across healthcare, finance and technology), and industry conferences.

- **Google Inc., Mountain View, CA, USA** Aug. 2012 — May 2019

  Staff Research Scientist, Google Brain (May 2016 — May 2019)

  Interpreting Machine Learning models

  - My research spanned the following objectives — (1) Increasing end-user transparency of ML models, (2) Evaluating robustness of ML models, and (3) Extracting human-intelligible rules from ML models (**refer Publications [3, 4, 12, 14, 15, 16, 17, 34], Patent [1]**).
  - Co-developed a method called "**Integrated Gradients**" for attributing a model's predictions to its input features **(refer Publication 17)**. This work has had **tremendous impact** with usage by 25+ product teams at Google, launch as part of the Cloud XAI product, **5000+** citations, and independent implementation in industry across several open-source ML frameworks.

  Question-Answering Models

  - Drove research on techniques for blending machine learning with traditional semantic parsing approaches to question-answering, and assessing robustness of question-answering systems (**refer Publications [15, 35, 36]**).

  Machine Learning Fairness

  - Developed techniques for addressing fairness concerns in text classification models. Launched as part of Youtube's brand safety system (**refer Publication [13]**).

  Senior Research Scientist, Security Research (Aug. 2012 — May 2016)

  Identity and Authorization

  - Designed and implemented the security model for the *Vanadium* application framework; see https://v.io. The model supports decentralized identities, mutual authentication and authorization, fine-grained delegation, and end-to-end encryption.(**refer Publications [5, 18, 19], Patents [2, 3, 4]**)
  - Developed the core technology for *Macaroons* — a flexible authorization credential for decentralized and controlled delegation of authority in distributed systems. Macaroons have seen widespread adoption inside and outside Google with open-source implementations in 9 different languages; see http://macaroons.io (**refer Publication [20], Patent [5]**)

Education

- **Stanford University**

  Ph.D. in Computer Science Jun. 2012

  Thesis title: "Sandboxing Untrusted JavaScript"

  Advisor: Prof. John C. Mitchell

  Google Ph.D. Fellow

  M.S. in Computer Science Jun. 2010

- **Indian Institute of Technology (IIT), Bombay**

  B. Tech in Computer Science and Engineering May 2007

Selected Awards and Honors

- **5 spot bonuses** at Google since 2023 for product launches across LLM factuality and ML explainability.
- **Outstanding paper award** for Publication [10], Cross Domain (CD) Conference for Machine Learning and Knowledge Extraction (CD-MAKE), Aug. 2020.
- **Outstanding paper award** for Publication [18], 21^{st} European Symposium on Research in Computer Security (ESORICS), Sep. 2016.
- 3^{rd} **prize** for Publication [22], **AT&T Best Applied Security Paper Award** competition, Nov. 2011.
- **Google PhD Fellowship in Language Security (2010-2012)**, Jun. 2010.
- Selected for **best papers from VMCAI 2009**, 10^{th} Int'l conference on Verification, Model checking and Abstract Interpretation, for Publication [29], Dec. 2009.
- **All India Rank 69** out of 180,000 students, IIT joint entrance examination, Jun. 2003.
- **Gold Medal**, Indian National Physics Olympiad, May 2003.
- **Gold Medal**, Indian National Chemistry Olympiad, May 2003.

Professional Activities

- **Teaching**:
  - **Graduate courses**:
    - **CS 328T: Trustworthy Machine Learning** (Fall 2023, Stanford University): Co-taught with Prof. John C. Mitchell and Prof. Anupam Datta.
    - **Distributed Authorization** (Summer 2016, FOSAD): Short course at International **summer school** on Foundations of Security, Analysis, and Design (FOSAD) at Bertinoro, Italy.
  - **Tutorials**: Co-organized tutorials on "Explainable AI" at KDD 2019, FAccT 2020, WWW 2020.
  - **Guest lectures**:
    - COMPSCI 282BR: Topics in Machine Learning, Harvard University, USA Mar. 2021
    - ECE 739: Security and Fairness of Deep Learning, CMU Silicon Valley, USA Feb. 2020
    - ECE 737: Engineering Safe Software Systems, CMU Silicon Valley, USA, Oct. 2019
    - CS 223: Advanced Computer Security, UC Santa Cruz, USA Oct. 2012
    - CS 258: Programming Language Theory, Stanford University, USA Mar. 2009
    - CS 242: Programming Languages, Stanford University, USA Oct. 2008
- **PhD thesis committee member**:
  - Pramod Kaushik Mudrakarta, The University of Chicago, 2019
  - John Sipple, George Washington University, 2023
- **Students mentored**: Andrew Bai (UCLA), Chih-Kuan Yeh (CMU), Susan Hao (UC Berkeley), Sahaj Garg (Stanford University), Pramod Kaushik Mudrakarta (The University of Chicago), Siddhartha Jayanti (Princeton University), Andres Erbsen (MIT).
- **Program committees:** ACM FAccT (2022), PLDI (ERC) 2019, ACM PLDI (ERC) 2014, ETAPS POST 2014, ACM PLAS 2013, HOTSPOT 2013.
- **Other committees:** DARPA ISAT study group on generative AI (2024), Dagstuhl seminars on Machine learning and Formal methods (2017), Scripting languages (2012).

Research Themes

- **Natural Language Processing [NLP]**: Analyzing and developing text models.
- **Interpreting Machine Learning models [IML]**: Interpreting the behavior of ML models.
- **Distributed Authorization [DA]**: Identity and Authorization in distributed systems.
- **Programming Language Security [PLS]**: Securing untrusted code.
- **Verification and Synthesis [VS]**: Verifying and synthesizing hardware, software, and hybrid systems.

Publications

Journals and Book Chapters:

1. [NLP] Clark Barrett, Brad Boyd, Elie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, Diyi Yang — *Identifying and Mitigating the Security Risks of Generative AI*, In: Proceedings of Foundations and Trends in Privacy and Security, 2023

2. [IML] Aya Abdelsalam Ismail, Sercan Ö. Arik, Jinsung Yoon, Ankur Taly, Soheil Feizi, Tomas Pfister — *Interpretable Mixture of Experts*, In: Proceedings of Transaction on Machine Learning Research (**TMLR**), 2023

3. [IML] Kevin McCloskey, Ankur Taly, Federico Monti, Michael P. Brenner, Lucy Colwell — *Using Attribution to Decode Dataset Bias in Neural Network Models for Chemistry*, In: Proceedings of National Academy of Science (**PNAS**) Preprint, 2019.

4. [IML] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jessie Smith, Arjun B. Sood, Greg S. Carrado, Lily Peng, Dale R. Webster — *Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy*, In: American Academy of Opthalmology (**AAO**), 2018.

5. [DA] Ankur Taly, Asim Shankar—*Distributed Authorization in Vanadium*, In: Lecture Notes on Foundations of Security, Analysis and Design (**FOSAD**), Springer, 2016.

6. [VS] Ankur Taly, Sumit Gulwani, Ashish Tiwari — *Synthesizing Switching Logic using Constraint Solving*, In: International Journal on Software Tools for Technology Transfer (**STTT**), Springer, 2011.

Conferences and Workshops:

7. [NLP, IML] Chih-Kuan Yeh, Ankur Taly, Mukund Sundararajan, Frederick Liu, Pradeep Ravikumar — *First is better than Last for Language Data Influence*, In: Neural Information Processing Systems (**NeurIPS**), 2022

8. [IML] David Watson, Limor Gultchin, Ankur Taly, Luciano Floridi — *Uncertainty in Artificial Intelligence* (**UAI**), 2021.

9. [IML] Ana Lucic, Madhulika Srikumar, Umang Bhatt, Alice Xiang, Ankur Taly, Q. Vera Liao, Maarten de Rijke — *A Multistakeholder Approach Towards Evaluating AI Transparency Mechanisms*, In: **CHI Workshop** on Operationalizing Human-Centered Perspectives in Explainable AI, 2020

10. [IML] Luke Merrick, Ankur Taly — *The Explanation Game: Explaining Machine Learning Models using Shapley Values*, In: Cross Domain conference on Machine Learning and Knowledge Extraction (**CD-MAKE**), 2020. (**award paper**)

11. [IML] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, Peter Eckersley — *Explainable Machine Learning in Deployment*, In: ACM Fairness, Accountability and Transparency (**FAccT**), 2020.

12. [IML, VS] Divya Gopinath, Hayes Converse, Corina Pasareanu, Ankur Taly — *Finding Contracts in Deep Neural Networks*, In: Automated Software Engineering (**ASE**), 2019.

13. [NLP, IML] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, Alex Beutel — *Counterfactual Fairness in Text Classification through Robustness*, In: ACM Artificial Intelligence, Ethics, and Society (**AIES**), 2019.

14. [IML] Mukund Sundararajan, Jinhua Xu, Ankur Taly, Rory Sayres, Amir Najmi — *Exploring Principled Visualization of Deep Network Attributions*, In: IUI Workshop on Explainable Smart Systems (**ExSS**), 2019.

15. [NLP, IML] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere — *Did the model understand the question?*, In: Annual Meeting of the Association for Computational Linguistics (**ACL**), 2018.

16. [IML] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Lily Peng, Dale R. Webster — *Assisted reads for diabetic retinopathy using a deep learning algorithm and integrated gradient explanation* (extended abstract), Annual meeting of the Association for Research in Vision and Ophtalmology (**ARVO**), 2018.

17. [IML] Mukund Sundararajan, Ankur Taly, Qiqi Yan — *Axiomatic Attribution for Deep Networks*, In: International Conference on Machine Learning (**ICML**), 2017. (**5000+ citations**)

18. [DA] David Wu, Ankur Taly, Asim Shankar, Dan Boneh — *Privacy, Discovery and Authentication for Internet of Things*, In: European Symposium on Research in Computer Security (**ESORICS**), 2016 (**award paper**).

19. [DA] Martin Abadi, Mike Burrows, Himabindu Pucha, Adam Sadovsky, Asim Shankar, Ankur Taly— *Distributed Authorization With Distributed Grammars*, In: Programming Languages with Applications to Biology and Security (**PLABS**), 2015.

20. [DA] Arnar Birgisson, Joe Politz, Ulfar Erlingsson, Ankur Taly, Michael Vrable, Mark Lentczner — *Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud*, In: Network and Distributed System Security Symposium (**NDSS**), 2014.

21. [VS] Patrice Godefroid, Ankur Taly — *Automated Synthesis of Symbolic Instruction Encodings from I/O Samples*, In: ACM Programming Language Design and Implementation (**PLDI**), 2012.

22. [PLS, VS] Ankur Taly, Ulfar Erlingsson, John C. Mitchell, Mark S. Miller, Jasvir Nagra — *Automated Analysis of Security-Critical JavaScript APIs*, In: IEEE Symposium on Security and Privacy (**S&P**), 2011 (**award paper**).

23. [VS] Ankur Taly, Ashish Tiwari — *Switching Logic Synthesis for Reachability*, In: ACM International Conference on Embedded Software (**EMSOFT**), 2010.

24. [PLS] Sergio Maffeis, John C. Mitchell, Ankur Taly — *Object Capabilities and Isolation of Untrusted Web Applications*, In: IEEE Symposium on Security and Privacy (**S&P**), 2010.

25. [VS] Ankur Taly, Ashish Tiwari - *Deductive Verification of Continuous Dynamical Systems*, In: IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (**FST&TCS**), 2009.

26. [LS] Sergio Maffeis, John C. Mitchell, Ankur Taly — *Isolating JavaScript with Filters, Rewriting, and Wrappers*, In: European Symposium on Research in Computer Security (**ESORICS**), 2009.

27. [PLS] Sergio Maffeis, Ankur Taly - *Language-Based Isolation of Untrusted JavaScript*, In: IEEE Symposium on Computer Security Foundations (**CSF**), 2009.

28. [PLS] Sergio Maffeis, John C. Mitchell, Ankur Taly — *Run-Time Enforcement of Secure JavaScript Subsets*, Web 2.0 Security and Privacy (**W2SP**) workshop, 2009.

29. [VS] Ankur Taly, Sumit Gulwani, Ashish Tiwari — *Synthesizing Switching Logic using Constraint Solving*, In: International Conference on Verification, Model Checking and Abstract Interpretation (**VMCAI**), 2009 (**selected as one of the best papers**).

30. [PLS] Sergio Maffeis, John C. Mitchell, Ankur Taly — *An Operational Semantics for JavaScript*, In: Asian Programming Languages Symposium (**APLAS**), 2008.

31. [VS] Stephane Gaubert, Eric Goubault, Ankur Taly, Sarah Zennou — *Static Analysis by Policy Iteration on Relational domains*, In: European Symposium on Programming (**ESOP**), 2007.

32. [VS] Sudeep Juvekar, Ankur Taly, Varun Kanade, Supratik Chakraborty — *Efficient Symbolic Reachability of Networks of Transition Systems*, In: General Motors Workshop on Next Generation Design and Verification Methodologies for Distributed Embedded Control Systems (**GMRD**), 2007.

Preprints and Technical Reports:

33. [NLP] Andrew Bai, Chih-Kuan Yeh, Cho-Jui Hsieh, Ankur Taly — *Which Pretrain Samples to Rehearse when Finetuning Pretrained Models?*, in submission, 2024

34. [IML] Mukund Sundararajan, Ankur Taly — *A Note about: Local Explanation Methods for Deep Neural Networks lack Sensitivity to Parameter Values*, Technical Report, on arxiv, 2018.

35. [NLP] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere — *It was the training data pruning too!*, Technical report, on arxiv, 2018.

36. [NLP] Kedar Dhamdhere, Kevin Mccurley, Mukund Sundararajan, Ankur Taly — *Abductive Matching in Question Answering*, Technical report, on arxiv, 2017.

37. [PLS] Ankur Taly - *Separation Logic and Mashup Isolation*, Technical report, Stanford University, 2010.

Theses:

38. [PLS] Ankur Taly - *Sandboxing Untrusted JavaScript*, Doctoral thesis, Stanford University, 2013.

39. [VS] Ankur Taly - *Efficient Guided Symbolic Reachability Analysis*, Bachelor's thesis, IIT Bombay, 2007.

40. [VS] Ankur Taly - *Automata on Infinite Inputs*, Junior thesis, IIT Bombay, 2006.

Patents

1. Amir Najmi, Ankur Taly, Jinhau Xu, Rory Sayres, Mukund Sundararajan — *Attribution Methodologies for Neural Networks Designed for Computer-Aided Diagnostic Processes*, 2019

2. Michael Burrows, Himabindu Pucha, Raja Daoud, Jatin Lodhia, Ankur Taly — *Signatures Of Updates Exchanged In A Binary Data Synchronization Protocol*, 2017.

3. Martin Abadi, Mike Burrows, Himabindu Pucha, Adam Sadovsky, Asim Shankar, Ankur Taly — *Authorization in a Distributed System Using Access Control Lists and Groups*, 2017.

4. Ankur Taly, Asim Shankar, Gautham Thambidorai, David Presotto — *Security model for identification and authentication in encrypted communications using delegate certificate chain bound to third party key*, 2016.

5. Ulfar Erlingsson, Ankur Taly, Michael Vrable, Mark Lentczner - *Macaroons: Methods and Systems of Generating and Using Authentication Credentials for Decentralized Authorization in the Cloud*, 2016.

Recent Invited Talks

• Google Factuality Summit, New York, USA	Feb. 2024
• Generative AI Risks Workshop, Mountain View, USA	Jun. 2023
• National Academies Workshop on AI Safety, USA	Sep. 2022
• Open Data Science Conference (ODSC East), Boston, USA	Apr. 2022
• AICTE Faculty Development Program, India	Dec. 2021
• Intuit Tech Con, Mountain View, USA	Sep. 2019
• O'Reilly Artificial Intelligence Conference, San Jose, USA	Sep. 2019
• Fiddler Explainable AI Summit (keynote), San Francisco, USA	Aug. 2019
• GFMI Model Risk Conference, New York, USA	Jun. 2019
• DREAM seminar, UC Berkeley, Berkeley, USA	Feb. 2019
• CSL seminar, SRI International, Menlo Park, USA	Dec. 2018
• Samsung AI Research, Mountain View, USA	Sep. 2018
• Dagstuhl Workshop on Machine Learning and Formal Methods, Schloss Dagstuhl, Germany	Aug. 2017
• CSL seminar, SRI International, Menlo Park, USA	Jun. 2017
• Bhabha Atomic Research Center, Mumbai, India	Dec. 2015
• Keybase.io, San Francisco, USA	Aug. 2015
• Vail Computer Elements Workshop, Vail, USA	Jun. 2015

Last updated: June 26, 2024