Umang Bhatt

Contact Information
umang.s.bhatt@gmail.com — Website
Citizenship: USA — Google Scholar

Interests
Trustworthy Machine Learning, Responsible AI, Human-Machine Collaboration, AI Policy

Academic Positions

Assistant Professor & Faculty Fellow, **New York University** — Oct 2023 – Present
Center for Data Science

Senior Research Associate, **The Alan Turing Institute** — Jan 2023 – Present
Project ELSA: European Lighthouse on Secure and Safe AI

Research Fellow, **Harvard University** — Jun 2022 – Mar 2023
Center for Research on Computation and Society

Education

University of Cambridge, Cambridge, UK
Ph.D. in Engineering — Sept 2019 – Nov 2023
Passed with No Corrections
Thesis: Trustworthy Machine Learning: From Algorithmic Transparency to Decision Support
Committee: Adrian Weller MBE (Advisor), Neil Lawrence, Eric Horvitz
Affiliations: Machine Learning Group, Computation and Biological Learning Lab

Carnegie Mellon University, Pittsburgh, PA
M.S. in Electrical and Computer Engineering — Aug 2017 – May 2019
B.S. in Electrical and Computer Engineering — Aug 2015 – May 2019

Select Fellowships and Awards

Center for Democracy & Technology Fellowship — 2024 – 2026
J.P. Morgan AI PhD Fellowship *(Declined)* — 2022 – 2023
The Alan Turing Institute Enrichment Studentship — 2021 – 2022
Mozilla Fellowship — 2020 – 2021
Partnership on AI Research Fellowship — 2019 – 2020
Leverhulme Center for the Future of Intelligence PhD Scholarship — 2019 – 2023
Fully funded by Google DeepMind and the Leverhulme Trust
Best Presentation Award, AAAI Spring Symposium on Interpretable AI for Well-Being — 2019
Lovett Family Endowed Scholarship, The Andrew Carnegie Society — 2019
Undergraduate Research Presentation Award, CMU — 2017
NSF I-Corps Site Award for research commercialization — 2017
H. F. McCullough Memorial Scholarship, CMU — 2016

Journal and Conference Publications

[1] **Learning Personalized Decision Support Policies**
AAAI Conference on Artificial Intelligence (AAAI) 2025 **(Oral Presentation)**
Umang Bhatt*, Valerie Chen*, Katherine Collins, Parameswaran Kamalaruban, Emma Kallina, Adrian Weller, Ameet Talwalkar

[2] **Building Machines that Learn and Think *with* People**
Nature Human Behavior, 2024.
Katherine Collins*, Ilia Sucholutsky*, ***Umang Bhatt****, Kartik Chandra*, Lionel Wong*, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua Tenenbaum, Thomas Griffiths

[3] **When Should Algorithms Resign? A Proposal for AI Governance**
IEEE Computer, 2024
Umang Bhatt*, Holli Sargeant*

[4] **Evaluating Language Models for Mathematics through Interactions**
Proceedings of the National Academy of Sciences (PNAS), 2024.
Katherine Collins, Albert Jiang, Simon Frieder, Lionel Wong, Miri Zilka, ***Umang Bhatt***, Thomas Lukasiewicz, Yuhuai Wu, Joshua Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller, Mateja Jamnik

[5] **Large Language Models Must Be Taught to Know What They Don't Know**
Neural Information Processing Systems (NeurIPS), 2024.
Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, ***Umang Bhatt***, Adrian Weller, Samuel Dooley, Micah Goldblum, Andrew Gordon Wilson

[6] **Comparing Abstraction in Humans and Large Language Models Using Multimodal Serial Reproduction**
46th Annual Conference of the Cognitive Science Society (CogSci) 2024 ***(Oral Presentation)***
Sreejan Kumar, Raja Marjieh, Byron Zhang, Declan Campbell, Michael Y. Hu, ***Umang Bhatt***, Brenden Lake, Thomas L. Griffiths

[7] **Perspectives on Incorporating Expert Feedback into Model Updates**
Patterns, 2023.
Valerie Chen*, ***Umang Bhatt***, Hoda Heidari, Adrian Weller, Ameet Talwalkar

[8] **Algorithmic Loafing and Mitigation Strategies in Human-AI Teams**
Computers in Human Behavior: Artificial Humans, 2023.
Isa Inuwa-Dutse, Alice Toniolo, Adrian Weller, ***Umang Bhatt***

[9] **Selective Concept Models: Permitting Stakeholder Customisation at Test-Time**
AAAI Conference on Human Computation and Crowdsourcing (HCOMP) 2023
Matthew Barker, Katherine Collins, Krishnamurthy Dvijotham, Adrian Weller, ***Umang Bhatt***

[10] **FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines**
ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO) 2023
Matthew Barker, Emma Kallina, Dhananjay Ashok, Katherine Collins, Ashley Casovan, Adrian Weller, Ameet Talwalkar, Valerie Chen, ***Umang Bhatt***

[11] **Human-in-the-Loop Mixup**
Uncertainty in Artificial Intelligence (UAI) 2023 **(Oral Presentation)**
Katherine Collins, ***Umang Bhatt***, Weiyang Liu, Vihari Piratla, Ilia Sucholutsky, Bradley Love, Adrian Weller

[12] **On the Informativeness of Supervision Signals**
Uncertainty in Artificial Intelligence (UAI) 2023
Ilia Sucholutsky, Ruairidh McLennan Battleday, Katherine Collins, Raja Marjieh, Joshua Peterson, Pulkit Singh, ***Umang Bhatt***, Nori Jacoby, Adrian Weller, Thomas Griffiths

[13] **Human Uncertainty in Concept-Based AI Systems**
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) 2023
Katherine Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, ***Umang Bhatt***, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, Krishnamurthy Dvijotham

[14] **Iterative Teaching by Data Hallucination**
International Conference on Artificial Intelligence and Statistics (AISTATS) 2023
Zeju Qiu, Weiyang Liu, Tim Xiao, Zhen Liu, Yucen Luo, ***Umang Bhatt***, Adrian Weller, Bernhard Schölkopf

[15] **Approximating Full Conformal Prediction at Scale via Influence Functions**
AAAI Conference on Artificial Intelligence (AAAI) 2023
Javier Abad Martinez, ***Umang Bhatt***, Adrian Weller, Giovanni Cherubin

[16] **Towards Robust Metrics for Concept Representation Evaluation**
AAAI Conference on Artificial Intelligence (AAAI) 2023
Mateo Zarlenga, Pietro Barbiero, Zohreh Shams, D. Kazhdan, ***Umang Bhatt***, Adrian Weller, Mateja Jamnik

[17] **How Transparency Modulates Trust in Artificial Intelligence**
Patterns, 2022.
John Zerilli, ***Umang Bhatt***, Adrian Weller

[18] **Uncertainty Quantification with Pre-trained Language Models: An Empirical Analysis**
Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022 (Findings)
Yuxin Xiao, Paul Pu Liang, ***Umang Bhatt***, Willie Neiswanger, Ruslan Salakhutdinov, L.P. Morency

[19] **Eliciting and Learning with Soft Labels from Every Annotator**
AAAI Conference on Human Computation and Crowdsourcing (HCOMP) 2022
Katherine Collins*, ***Umang Bhatt***, Adrian Weller

[20] **On the Utility of Prediction Sets in Human-AI Teams**
International Joint Conference on Artificial Intelligence (IJCAI) 2022 **(Oral Presentation)**
Varun Babbar, ***Umang Bhatt***, Adrian Weller

[21] **Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates**
AAAI Conference on Artificial Intelligence (AAAI) 2022
Dan Ley, ***Umang Bhatt***, Adrian Weller

[22] **On the Fairness of Causal Algorithmic Recourse**
AAAI Conference on Artificial Intelligence (AAAI) 2022 **(Oral Presentation)**
Julius von Kügelgen, Amir Karimi, ***Umang Bhatt***, Isabel Valera, Adrian Weller, Bernhard Schölkopf

[23] **Uncertainty as a Form of Transparency: Measuring and Communicating Uncertainty**
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) 2021
Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, Alice Xiang

[24] **Getting a CLUE: A Method for Explaining Uncertainty Estimates**
International Conference on Learning Representations (ICLR) 2021 **(Oral Presentation)**
Javier Antorán, ***Umang Bhatt***, Tameem Adel, Adrian Weller, José Miguel Hernández-Lobato

[25] **FIMAP: Feature Importance by Minimal Adversarial Perturbation**
AAAI Conference on Artificial Intelligence (AAAI) 2021
Matt Chapman-Rounds, ***Umang Bhatt***, Erik Pazos, Marc-Andre Schulz, Kostas Georgatzis

[26] **Evaluating and Aggregating Feature-based Explanations**
International Joint Conference on Artificial Intelligence (IJCAI) 2020
Umang Bhatt, Adrian Weller, José M.F. Moura

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

[28] **You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods**
European Conference on Artificial Intelligence (ECAI) 2020
Botty Dimanov, ***Umang Bhatt***, Mateja Jamnik, Adrian Weller

[29] **On Network Science and Mutual Information for Explaining Deep Neural Networks**
IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020
Brian Davis*, ***Umang Bhatt***, Kartikeya Bhardwaj*, Radu Marculescu, José M.F. Moura

Book Chapters

[30] **Trust in Artificial Intelligence: Clinicians are Essential**
Healthcare Information Technology for Cardiovascular Medicine, 2021
Umang Bhatt, Zohreh Shams

Teaching Experience

Course Instructor, *New York University*

- Responsible Data Science (DS-UA 202) — Spring 2025
  Co-Taught with Jonathan Colner
- Responsible Data Science (DS-UA 202) — Spring 2024
  Topics: Fairness, Accountability, Transparency, and Privacy

Teaching Assistant (Supervisor/Grader), *University of Cambridge*

- *Inference* (3F8) for J.M. Hernández-Lobato and David Krueger — Lent 2023

- *Inference* (3F8) for Richard Turner and David Krueger — Lent 2022
- *Probabilistic ML* (4F13) for Zoubin Gharamani and J.M. Hernández-Lobato — Michaelmas 2020

Teaching Assistant, *Carnegie Mellon University*

- *Machine Learning for Engineers - Masters* (18-661) for Gauri Joshi — Spring 2019
- *Machine Learning - PhD* (10-701) for Ziv Bar-Joseph and Pradeep Ravikumar — Fall 2018
- *Practical Data Science* (15-688) for Zico Kolter — Spring 2018

Supervision

Advisor, *Alan Turing Institute*

- Mackenzie Jorgensen, Postdoctoral Research Associate — Nov 2024 – Present
- Kendall Brogle, Research Assistant — Sept 2024 – Present

Advisor, *New York University*

- Lujain Ibrahim, Visiting PhD Student — Nov 2024 – Present
- Mihir Upadhyay, MS in Data Science — Aug 2024 – Present
- Elaf Almahmoud, MS in Computer Science — Jun 2024 – Present
- Ghina Al Shdaifat, MS in Data Science — Apr 2024 – Present
- Arina Shah, BA in Philosophy — Feb 2024 – Present
- Suha Memon, BS in Computer Science *(Next: MS at Penn)* — Jan 2024 – June 2024

Thesis Co-Supervisor/Mentor, *University of Cambridge*

- Matthew Barker, MEng in Information Engineering *(Next: Trustwise)* — May 2022 – Jun 2023
- Vivek Palaniappan, MEng in Information Engineering *(Next: Citadel)* — May 2022 – Jun 2023
- Katherine Collins, MPhil in Machine Learning *(Next: PhD at Cambridge)* — Nov 2021 – Sept 2022
  Departmental Distinction Awardee
- Varun Babbar, MEng in Information Engineering *(Next: PhD at Duke)* — May 2021 – Jun 2022
  2022 Engineering Division F Thesis Prize Winner
- Javier Abad Martinez, Research Assistant *(Next: PhD at ETH Zurich)* — Nov 2021 – Feb 2022
- Dan Ley, MEng in Information Engineering *(Next: PhD at Harvard)* — May 2020 – Aug 2021
  2021 Engineering Division F Thesis Prize Winner

Select Invited Talks

- TU Dortmund, *RC Trustworthy Data Science Seminar* — Jul 2024
- Google, *Responsible AI Talk Series* — Jul 2024
- Finnish Center for Artificial Intelligence, *ELLIS HCML Workshop* — Jun 2024
- KTH Royal Institute of Technology, *RPL Summer School* — Jun 2024
- US Department of Defense, *CDAO Seminar* — Feb 2024
- KAUST, *Rising Stars in AI Symposium* — Feb 2024
- Instituto dos Advogados de São Paulo - IASP, *Seminar* — Dec 2023
- House of Lords Select Committee on AI in Weapon Systems — Jun 2023
- Indian Institute of Technology Bombay, *C-MInDS Seminar* — Jan 2023
- University of Chicago, *Chicago Human+AI Lab* — Nov 2022
- Massachusetts Institute of Technology, *AI Ethics and Policy Group* — Oct 2022
- Birkbeck, University of London, *Workshop on Human Behavioral Aspects of XAI* — Sept 2022
- DeepMind, *Ethics Research Team* — July 2022
- University of Bristol, *REPHRAIN Masterclass* — June 2022
- University of Manchester, *Advances in Data Science and AI Conference* — June 2022
- Leverhulme Centre for the Future of Intelligence, *Seminar Series* — June 2022
- Ada Lovelace Institute, *Brown Bag Lunch Seminar* — June 2022
- Von Hügel Institute, *Research Salon on Artificial Intelligence* — June 2022
- Pennsylvania State University, *Young Achievers Symposium* — Apr 2022
- UK Ministry of Defense, *AI Safety Workshop* — Mar 2022
- Vanguard Health, *MedTech Insight Podcast* — June 2021
- Technical University of Denmark, *Trustworthiness and Interpretability in ML Seminar* — Apr 2021
- Harvard SEAS, *Guest Lecture for COMPSCI 282BR: Explainability in ML* — Apr 2021
- Imperial College, *Explainable AI Seminar* — Feb 2021
- Robust and Responsible AI Developers *(Keynote)* — July 2020
- ICML Workshop on Extending Explainable AI *(Keynote)* — July 2020

- Mozilla All-Hands Meeting *(Keynote)* June 2020
- QuantumBlack (McKinsey), *AI Seminar* May 2020
- Fiddler Labs, *Explainable AI Seminar* May 2019

Professional Service

Finance Chair, ACM Conference on Fairness, Accountability, and Transparency (FAccT) 2024

Associate Chair, International Conference on Machine Learning (ICML) 2024

Co-Organizer

- DALI Meeting: Communication of Uncertainty Workshop Apr 2025
- Deep Learning Indaba: Responsible AI Practical Workshop Sept 2024
- ELSA Workshop on Generative AI and Creative Arts Mar 2024
- Deep Learning Indaba: Responsible AI Practical Workshop Sept 2023
- ICML Workshop on Human-Machine Collaboration and Teaming July 2022
- ELLIS Workshop on Human-Centric Machine Learning May 2021
- ICML Workshop on Human Interpretability in Machine Learning July 2020
- IBM + Partnership on AI Workshop on Explainable AI Feb 2020

Program Committee (Reviewer)

- 2025: AISTATS, FAccT (Area Chair), ICML, UAI
- 2024: AIES, FAccT, NeurIPS
- 2023: FAccT, ICLR, ICML, IJCAI, NeurIPS, UIST, WebConf
- 2022: AAAI, AISTATS, FAccT, ICLR, ICML, NeurIPS, UAI
- 2021: AAAI, AISTATS, FAccT, ICAIF, ICLR, ICML, KDD, NeurIPS, UAI
- 2020: ICAIF, NeurIPS

Reviewer: ACM Computing Surveys, ACM Transactions on Computer-Human Interaction (TOCHI), ACM Transactions on Interactive Intelligent Systems (TiiS), Artificial Intelligence Journal (AIJ), Harvard Data Science Review (HDSR), IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Journal of Artificial Intelligence Research (JAIR), Journal of Machine Learning Research (JMLR), Transactions on Machine Learning Research (TMLR), Management Science

Professional Experience

Scientific Advisor, **Multiple Startups** Nov 2024 – Present
- Starlab (Barcelona, Spain), Reflective Works (New York, NY)

Advisor, **Slalom**, New York, NY Nov 2023 – Present
- Responsible AI Client Advisory Board

Advisory Panel, **Bourne-Epsom Protocol**, London, UK Jul 2023 – Present
- Expert guidance on the technical aspects of AI in education

Advisor, **OECD**, Paris, FR Apr 2023 – Present
- Expert Group on AI Risk & Accountability

Advisory Chief Scientist, **Trustwise**, London, UK Jan 2023 – Present
- Building safety guardrails for large language models

Subject Matter Expert, **Accenture**, New York, NY Jan 2023 – Apr 2024
- Advising on Responsible AI strategy

Advisor, **Responsible AI Institute**, Austin, TX Oct 2021 – Jan 2023
- Building a scalable certification program for AI systems

Advisor, **Credo AI**, Palo Alto, CA Jun 2020 – Jun 2021
- Scoped an AI governance and auditing platform; backed by AI Fund

Research Assistant, **Carnegie Mellon University**, Pittsburgh, PA Jan 2017 – Sept 2019
- Advisors: José M.F. Moura (ECE), Pradeep Ravikumar (MLD), and Zico Kolter (CSD)

Student Fellow, **.406 Ventures**, Boston, MA July 2017 – Jun 2019
- Sourced startups and performed first-round due diligence on ventures

Intern, **Microsoft**, Redmond, WA May 2018 – Aug 2018
- Project: explainable conversational agents for technical hardware documentation

Co-Founder, **Percepsense**, Pittsburgh, PA Jan 2017 – May 2018
- Built products to harvest vehicular telematics data; pipeline acquired by ***Honda Motors***