Shomir Wilson

The Pennsylvania State University
College of Information Sciences and Technology
E310 Westgate Building
University Park, PA 16802

shomir@psu.edu
https://shomir.net

EDUCATION

University of Maryland	2011 **Ph.D.**, Computer Science; 2008 **M.S.**, Computer Science
Virginia Tech	2005 **B.S.**, Computer Science, **B.S.**, Mathematics, **B.A.**, Philosophy
U. of Mary Washington	1999 part-time, non-degree seeking

EXPERIENCE

Primary Affiliations

Pennsylvania State University College of Information Sciences and Technology (University Park, PA)	8/2018 – now	**Assistant Professor** Director, Human Language Technologies Lab Graduate Faculty, College of IST Graduate Faculty, Social Data Analytics Affiliate, Center for Socially Responsible Artificial Intelligence Affiliate, Institute for Computational and Data Sciences Affiliate, Social Science Research Institute
University of Cincinnati EECS Department (Cincinnati, OH)	8/2016 – 8/2018	**Assistant Professor** Director, Human Language Technologies Lab Member, Institute for Analytics Innovation
Carnegie Mellon University School of Computer Science (Pittsburgh, PA)	8/2015 – 8/2016	**Project Scientist** *Supervisor: Norman Sadeh*
	1/2015 – 5/2015	**Lecturer** (Natural Language Processing)
	8/2014 – 7/2015	**NSF International Research Fellow** *Supervisor: Alan W Black*
	9/2011 – 7/2013	**Postdoctoral Fellow** *Supervisor: Norman Sadeh*

Visiting Affiliations

University of Edinburgh School of Informatics (Edinburgh, United Kingdom)	8/2013 – 7/2014	**NSF International Research Fellow** *Host: Jon Oberlander*
Nat'l Univ. of Singapore School of Computing (Kent Ridge, Singapore)	6/2010 – 8/2010	**NSF EAPSI Fellow** *Host: Min-Yen Kan*
Macquarie University School of Computing (Sydney, Australia)	6/2009 – 8/2009	**NSF EAPSI Fellow** *Host: Robert Dale*

GRANTS AND SPONSORSHIP

Financial Support: Sources External to the University

1. **National Science Foundation**: PI on #2237574 "CAREER: Large-Scale Exploration and Interpretation of Consumer-Oriented Legal Documents". **$556,397 to Penn State** for 2023-08-01 – 2028-07-31.
2. **National Science Foundation**: Co-PI on #2247723 "SaTC: CORE: Small: Toward Privacy Equity through Contextual Understanding of Self-Disclosure". Other senior personnel: Sarah Rajtmajer (PI). **$599,851 to Penn State** for 2023-06-01 – 2026-05-31.
3. **National Science Foundation**: Lead PI on #2105736 "Collaborative Research: SaTC: CORE: Medium: A Large-Scale, Longitudinal Resource to Advance Technical and Legal Understanding of Textual Privacy Information". Other senior personnel: Lee Giles (Co-PI), Florian Schaub (PI at University of Michigan), Gabriela Zanfir-Fortuna (PI at Future of Privacy Forum). **$699,744 to Penn State and $1,198,820 overall + $21,000 REU supplement to Penn State** for 2021-07-01 – 2024-06-30.
4. **National Institutes of Health**: Co-I (subawardee) on #1R01MD015064-01A1 "Primed to (re)act: Can changes in procedural language prevent adverse events between police and minority male youth?". Other senior personnel: Margaret Beale Spencer (PI at University of Chicago), Christopher Graziul (PI at University of Chicago), Karen Livescu (Co-I at Toyota Technological Institute at Chicago), Lisa Thurau (Co-I at Strategies for Youth). **$228,054 to Penn State** for 2021-01-27 – 2023-12-30.
5. **National Science Foundation**: PI on #1914444 "SaTC: CORE: Medium: Collaborative: Automatically Answering People's Privacy Questions". Other senior personnel: Norman Sadeh (Lead PI at Carnegie Mellon University), Joel Reidenberg (PI at Fordham University), Tom Norton (Senior Personnel at Fordham University). **$437,436 + $14,000 REU supplement to Penn State** for 2019-07-15 – 2022-12-31.
6. **National Science Foundation**: Subawardee on #1330596 "TWC SBE: Option: Frontier: Collaborative: Towards Effective Web Privacy Notice and Choice: A Multi-Disciplinary Prospective". Other senior personnel: Norman Sadeh (Lead PI at Carnegie Mellon University),

et al. **$76,249 to University of Cincinnati** for 2016-09-01 – 2018-07-31; **$13,792 to Penn State** for 2018-09-01 – 2019-07-31.

7. **National Science Foundation**: PI on #1159236 "IRFP: Metalanguage Identification for Interactive Language Technologies". **$98,100** fellowship for 2013-08-01 – 2015-03-31.
8. **National Science Foundation**: PI on #1015666 "EAPSI: Parsing Metalanguage and the Use-Mention Distinction". **$5,000 + travel support** fellowship for 2010-06-13 – 2010-08-06.
9. **National Science Foundation**: PI on #0914091 "EAPSI: Distinguishing Use and Mention in Natural Language". **$5,000 + travel support** fellowship for 2009-06-22 – 2009-08-16.

Financial Support: Sources Internal to the University

1. **Penn State Center for Socially Responsible Artificial Intelligence**: Co-PI on "Understanding the Prevalence of Drinking Water Service Disruption through Large-Scale Analysis of News Articles and Social Media". Other senior personnel: Christine Kirchhoff (Lead PI). **$25,000** for 2022.
2. **Penn State Center for Social Data Analytics**: PI on "Exploring the Effects of Socioeconomic Status on Privacy Behaviors in an Online Social Network". Other senior personnel: Sarah Rajtmajer (Co-PI). **$25,000** for 2020.
3. **Penn State Center for Cybersecurity Research and Education**: PI on "Automatic Detection of Malicious Emails Using Discourse Analysis". Other senior personnel: Susan Strauss (Co-PI). **$7,500** for 2019 – 2021.
4. **University of Cincinnati**: PI on College of Engineering and Applied Sciences Faculty Development Grant. **$2,350** for 2017.

Additional Support

1. **Pennsylvania Space Grant Consortium**: **Scholarship support to place undergraduate researchers in my lab** for 2019 – 2022.
2. **Penn State College of IST**: Lead PI on "Web-Scale Search and Analysis of Privacy Policies". Other senior personnel: Lee Giles (Co-PI). **2 years GRA support** for 2019-08-26 – 2021-08-20.
3. **Ohio Supercomputer Center**: **9,000 Resource Units for research and classroom use** for 2016 – 2018.

PUBLICITY AND RECOGNITION

Press Interviews and Mentions

Recorded interview for TV news segment, "AI deepfakes becoming more lifelike, more misleading", WPMT, 2023-08-07. [hyperlink]

Recorded interview for TV news segment, "How AI is changing the industry and what it means for Pa. businesses", WPMT, 2023-08-06. [hyperlink]

Recorded interview for TV news segment, "How the AI revolution could impact education", WPMT, 2023-07-21. [hyperlink]

Quoted in "Going Phishing on Campus", Inside Higher Ed, 2023-07-18. [hyperlink]

Quoted in "AI Will Heighten Cybersecurity Risks for RIAs", Wealth Management, 2023-07-12. [hyperlink]

Recorded interview for TV news segment "Artificial intelligence programs are causing concern for educators", WTAJ, 2023-05-17. [hyperlink]

Recorded interview for TV news segment about artificial intelligence being used by swatting perpetrators, Centre County Report, 2023-04-21. [hyperlink]

Live interview about the potential impact of ChatGPT on online dating, NBC News NOW, 2023-03-17. [hyperlink]

Quoted in "How bias creeps into the AI designed to detect toxicity", VentureBeat. 2021-12-09. [hyperlink]

Research featured in "Textos en latín y longitudes inasumibles: las revelaciones de un buscador de políticas de privacidad" (*Latin texts and unaffordable lengths: The revelations of a privacy policy search engine*), El País. 2021-11-03. [hyperlink]

Quoted in "Will Google ever lose its throne as king of search? Here are its main contenders", Digital Trends. August 15, 2021. [hyperlink]

Recorded interview for TV news segment about FaceID privacy, WCPO Cincinnati. November 14, 2017.

National Science Foundation News from the Field: "Carnegie Mellon Researchers Create an AI to Help Us Make Sense of Privacy Policies", March 1, 2018. [hyperlink]

National Science Foundation SEE Innovation Research Highlight: "Refining a Computer's Understanding of Language", 2012.

Publicity by the University

iConnect Magazine: "Understanding the fine print", Winter 2023. [hyperlink]

Penn State News: "Positive triggering method reduces nationality bias in large text generators", 2023-04-25. [hyperlink]

Penn State News: "What is ChatGPT and what can it be used for?" 2023-04-21. [hyperlink]

Penn State News: "IST assistant professor Shomir Wilson receives NSF CAREER Award", 2023-04-21. [hyperlink]

Penn State News: "Email scammers tailor methods to target universities", 2023-04-20. [hyperlink]

Penn State News: "Center for Socially Responsible AI awards seed funding to 6 projects", 2023-01-27. [hyperlink]

Penn State News: "Researchers propose methods for automatic detection of doxing", 2022-12-09. [hyperlink]

Penn State News: "AI language models show bias against people with disabilities, study finds", 2022-10-13. [hyperlink]

iConnect Magazine: "Studying Adverse Police Interactions", Summer 2021. [hyperlink]

Penn State News: "Women and lower-education users more likely to tweet personal information", 2021-07-07. [hyperlink]

ICDS News: "$1.2 million NSF grant to create search engine for online privacy research", 2021-06-24. [hyperlink]

Penn State News: "Study of police language aims to find patterns that may lead to tragic outcomes", 2021-05-25. [hyperlink]

Penn State News: "What if opting out of data collection were easy?", 2021-01-14. [hyperlink]

Penn State News: "IST students and faculty 'build together' at Grace Hopper Celebration", October 22, 2020. [hyperlink]

iConnect Magazine: "Simplifying Privacy", Fall 2019. [hyperlink]

Penn State News: "Research Aims to Automatically Answer User Questions on Online Privacy Policies", July 23, 2019. [hyperlink]

Penn State News: "College of IST Awards Eight Seed Grants for Research Projects", June 13, 2019. [hyperlink]

Penn State News: "Penn State's Leadership in Artificial Intelligence Research", April 18, 2019. [hyperlink]

Penn State News: "Outside of IST: Faculty and Staff Making a Difference", October 17, 2018. [hyperlink]

Awards and Honors

Best Short Paper at the Third Workshop on Trustworthy Natural Language Processing (TrustNLP), 2023.

Graduation Speaker, Virginia Tech Department of Philosophy, May 18, 2019.

Best Paper Finalist at the 25th World Wide Web Conference (WWW), 2016.

University of Maryland International Conference Student Support Award, 2011.

University of Maryland Block Grant Fellowship, 2005–2007.

Virginia Tech Outstanding Senior in Computer Science, 2005.

Virginia Tech William H. Williams Senior Prize in Philosophy, 2004.

SERVICE

Service items with future dates are confirmed commitments.

Peer Review

Editorial Board, *Journal of Intelligent Information Systems*, 2018-present.

Standing Reviewer, *Computational Linguistics*, 2020-present.

Standing Reviewer, ACL Rolling Review, 2021-present.

Program Committee and/or Reviewer: AAAI 2024, IJCNLP-AACL 2023, ACL 2023, EACL 2023, EMNLP 2022, PrivateNLP 2022, ACL 2022, AAAI 2022, MISQ (2021), PrivateNLP 2021, NAACL 2021, EMNLP 2020, IMWUT 2020, ACL 2020, PrivateNLP 2020, AAAI 2019, EMNLP

2019, NAACL 2019, LREC TA-COS Workshop 2018, ICDCIT 2017, IEEE CIC PiCSoC Workshop 2016.

Grant Proposal Review for the Research Community: Marsden Fund of the Royal Society of New Zealand (2023), Swiss National Science Foundation (2020, 2019), National Science Foundation (2018, 2017, 2016), Portuguese Foundation of Science and Technology (FCT) (2012).

Grant Proposal Review for the Academic Unit and University: Institute for Cyber Science (2023), Center for Socially Responsible Artificial Intelligence (2023, 2022), College of IST Seed Grants (2023, 2021, 2020), Center for Social Data Analytics (2020).

Research Community

Lead Organizer, Birds-of-a-Feather Session: "NLP on Legal Text", EACL 2023.

Lead Organizer, NSF SaTC PI Meeting Breakout Session (with Athina Markopoulou): "Privacy, Policy, and People", 2022.

Student Research Workshop Mentor, NAACL 2022.

Co-Lead for Group Mentoring Session, ACL 2020.

Student Mentorship Program Participant, ACL 2019.

Lead Organizer, NSF SaTC PI Meeting Breakout Session (with Norman Sadeh): "Making Privacy Understandable for Internet Users", 2019.

Lead Organizer, the AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies, 2019. [hyperlink]

Session Chair, Social Applications II, EMNLP 2018.

Lead Organizer, the AAAI Fall Symposium on Privacy and Language Technologies, 2016. [hyperlink]

Session Chair, Logic Programming and Knowledge Engineering at AAAI 2008.

Department or College

Co-Organizer for IST-EECS Natural Language Processing Colloquium Series, Fall 2019 - present.

College of IST Awards Committee: Chair for Fall 2023 - Spring 2024, Member for Fall 2022 - Spring 2023.

Selection Committee for IST's Westin Scholars Award: Founder in 2021; Chair 2021 - 2022; Member 2023 - present.

Course Committee Chair for Integration of Security and Privacy (SRA 472), 2021 - present.

Course Committee Chair for First-Year Seminar for Data Science Majors (PSU 17), 2021 - present.

Faculty Annual Review Committee Member, 2023.

College of IST Security/Privacy Faculty Hiring Committee Member: 2022-2023, 2019 - 2020, 2018 - 2019.

Committee Member for IST PhD Qualifying Exams, 2022, 2021 (committee chair), 2019.

College of IST Graduate Recruitment Committee Member, Fall 2020 - Spring 2022.

College Representative to the NSF Workshop on Departmental Plans for Broadening Participation in Computer Science, 2019 and 2022.

College of IST Committee to Assemble Proposal for an Undergraduate Degree in Artificial Intelligence and Informatics, 2019.

IST Senior Exit Survey Committee, 2019.

College of Engineering and Applied Sciences eLearning Task Force, 2018.

Member of the EECS Graduate Admissions Committee, 2017 - 2018.

Faculty Marshal for the College of Engineering and Applied Sciences: May 2018, May 2017.

Graduate Student Representative to the Computer Science Department Council, 2008 - 2009.

University

University Faculty Senate, Fall 2022 - Spring 2026.

University Planning Committee (via Faculty Senate), Fall 2022 - present.

Steering Committee, Center for Socially Responsible Artificial Intelligence, Fall 2022 - present.

Interviewer for the Penn State Millennium Scholars Program: 2023, 2022, 2021, 2020, 2019.

Advisory Board for DACTR (`https://dactrgroup.com/`) in the Nittany AI Challenge, 2020.

Participant in the Penn State – University of Auckland Research Collaboration Development Workshop, 2019.

Outreach

Grand Awards Judge, International Science and Engineering Fair: 2021, 2018, 2015, and 2012.

Judge, Pittsburgh Regional Science and Engineering Fair, 2016.

Judge, Pittsburgh Regional FIRST Lego League, 2013 and 2011.

TEACHING

Semester Courses

Integration of Privacy and Security (SRA 472): Fall 2023, Fall 2022, Spring 2022, Spring 2021: Penn State University.

Computational Foundations of Informatics (IST 510): Spring 2023, Spring 2022: Penn State University.

First-Year Seminar for Data Science Majors (PSU 17): Spring 2023, Fall 2021, Fall 2020: Penn State University.

Honors Introduction to Information, People, and Technology (IST 110H): Fall 2020, Fall 2019: Penn State University.

Natural Language Processing for Sentiment, Semantics, and Discourse (IST 597):

Spring 2020, Spring 2019: Penn State University.

Advanced Topics in Natural Language Processing (CS 7052): Spring 2018, Spring 2017: University of Cincinnati.

Natural Language Processing (CS 5134/6034; 11-411/11-611): Fall 2017: University of Cincinnati; Spring 2015: Carnegie Mellon University (co-taught at CMU with Chris Dyer and Alan W Black).

Introduction to Low-Level Programming Concepts (CMSC 212): (Teaching Assistant) Spring 2006, Fall 2005: University of Maryland.

Condensed Courses

Natural Language Processing Methods and Applications: Workshop for faculty from Ming Chi University of Taiwan, July 11-13, 2018, University of Cincinnati (Cincinnati, Ohio).

Ethics of Artificial Intelligence: March 11-15, 2018, Future University (Cairo, Egypt).

RESEARCH ADVISEES AND THESIS COMMITTEES

Current Ph.D. Advisees: Younes Karimi, Shahriar Shayesteh, Mukund Srinath, Pranav Venkit, Tianyang Zhao
Current Undergraduate Researchers: Grace Ciambrone, Juan Larenas

Completed graduate advisees are listed below with their graduation placements, when available.
Completed M.S. Advisees: Sonu Gupta (2022), Duo Pan (2021, Associate Application Developer at ADP), Soundarya Sundareswara (2021, Software Engineer at Apple), Abhijith Athreya (2018, Chief Engineer at Samsung R&D), Baradwaj Aryasomayajula (2018, Software Developer at Incedo)
Completed Undergraduate Honors Thesis Advisee: Kathryn Frankenberg
Past Undergraduate Researchers: Avi Bewtra, Minh Doan, Miranda Goodman, Matthew Ihlenfield, Samantha Kenny, Ananya Nijhawan, Nora O'Toole, Ellen Poplavska, Benjamin Siri, Josephine Soddano, Eesha Srivatsavaya, Kaitlyn Wassel

Ph.D. Thesis Committee Memberships: Chase Bloch (2023), Kaixuan Zhang (2021), Maha Aljohani (2018)
M.S. Thesis Committee Memberships: Xi Lu (2020), Raphael Rodriguez (2020), Sampurna Ravindranathan (2018), Abhro Mondal (2017)

PAPERS

Papers with future publication dates are accepted to appear in their respective venues.

Peer-Reviewed Conference Proceedings

1. Privacy Lost and Found: An Investigation at Scale of Web Privacy Policy Availability. Mukund Srinath, Soundarya Nurani Sundareswara, Pranav Venkit, C. Lee Giles, and Shomir Wilson. In *Proceedings of the 23rd ACM Symposium on Document Engineering* (**DocEng**), 2023. **Best Student Paper Award.**

2. Privacy Now or Never: Large-Scale Extraction and Analysis of Dates in Privacy Policy Text. Mukund Srinath, Lee Matheson, Pranav Venkit, Gabriela Zanfir-Fortuna, Florian Schaub, C. Lee Giles and Shomir Wilson. In *Proceedings of the 23rd ACM Symposium on Document Engineering* (**DocEng**), 2023.
3. Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles. Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. In *Proceedings of the Sixth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (**AIES**), 2023.
4. Creation and Analysis of a Corpus of Scam Emails Targeting Universities. Grace Ciambrone and Shomir Wilson. In *Companion Proceedings of the ACM Web Conference* (**WebConf**), 2023.
5. Nationality Bias in Text Generation. Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Kenneth Huang, and Shomir Wilson. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (**EACL**), 2023.
6. A Study of Implicit Language Model Bias against People with Disabilities. Pranav Venkit, Mukund Srinath, and Shomir Wilson. In *Proceedings of the 29th International Conference on Computational Linguistics* (**COLING**), 2022.
7. STAPI: An Automatic Scraper for Extracting Iterative Title-Text Structure from Web Documents. Nan Zhang, Shomir Wilson, and Prasenjit Mitra. In *Proceedings of the 13th Language Resources and Evaluation Conference* (**LREC**), 2022.
8. A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus. Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan O. Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. In *Proceedings of the 13th Language Resources and Evaluation Conference* (**LREC**), 2022.
9. Automated Detection of Doxing on Twitter. Younes Karimi, Anna Squicciarini, and Shomir Wilson. In *Proceedings of the 25th ACM Conference on Computer-Supported Cooperative Work And Social Computing* (**CSCW**), 2022.
10. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. Mukund Srinath, Shomir Wilson, and C. Lee Giles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (**ACL**), 2021.
11. Breaking Down Walls of Text: How Can NLP Benefit Consumer Privacy? Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (**ACL**), 2021.
12. PrivaSeer: A Privacy Policy Search Engine. Mukund Srinath, Soundarya Nurani Sundareswara, C. Lee Giles, and Shomir Wilson. In *Proceedings of the 21st International Conference on Web Engineering* (**ICWE**), 2021.
13. A Large-Scale Exploration of Terms of Service Documents on the Web. Soundarya Nurani Sundareswara, Mukund Srinath, Shomir Wilson and C. Lee Giles. In Proceedings of the 21st ACM Symposium on Document Engineering (**DocEng**), 2021.
14. From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme. Ellen Poplavska, Thomas B. Norton, Shomir Wilson, and Norman Sadeh. In

Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems (**JURIX**), December 9-11, 2020.

15. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub and Norman Sadeh. In *Proceedings of The Web Conference* (**WWW**), April 20-24, 2020.

16. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. Abhilasha Ravichander, Alan Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), November 3-7, 2019.

17. Vaccine: Obfuscating Access Pattern Against File-Injection Attacks. Hao Liu, Boyang Wang, Nan Niu, Shomir Wilson, and Xuetao Wei. In *Proceedings of the IEEE Conference on Communications and Network Security* (**CNS**), June 10-12, 2019.

18. Supervised and Unsupervised Methods for Robust Separation of Section Titles and Prose Text in Web Documents. Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), Brussels, Belgium, November 2018.

19. Identifying the provision of choices in privacy policy text. Kanthashree Sathyendra, Shomir Wilson, Florian Schaub, Norman Sadeh, and Sebastian Zimmeck. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), Copenhagen, Denmark, September 2017.

20. Automated analysis of privacy requirements for mobile apps. Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. In *Proceedings of the Network and Distributed System Security Symposium* (**NDSS**), San Diego, California, March 2017.

21. The creation and analysis of a website privacy policy corpus. Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (**ACL**), Berlin, Germany, August 2016.

22. Crowdsourcing annotations of websites' privacy policies: Can it really work? Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah Smith and Frederick Liu. In *Proceedings of the 25th International World Wide Web Conference* (**WWW**), Montreal, Canada, April 2016. **Best Paper Finalist.**

23. This table is different: A WordNet-based approach to identifying references to document entities. Shomir Wilson, Alan W Black, and Jon Oberlander. In *Proceedings of the 8th International Global WordNet Conference* (**GWC**), Bucharest, Romania, January 2016.

24. Identifying relevant text fragments to help crowdsource privacy policy annotations. Rohan Ramanath, Florian Schaub, Shomir Wilson, Fei Liu, Norman Sadeh, and Noah Smith. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing* (**HCOMP**), works-in-progress track, Pittsburgh, PA, November 2014.

25. Determiner-established deixis to communicative artifacts in pedagogical text. Shomir Wilson and Jon Oberlander. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (**ACL**), Baltimore, MD, June 2014.
26. Toward automatic processing of English metalanguage. Shomir Wilson. In *Proceedings of the 6th International Joint Conference on Natural Language Processing* (**IJCNLP**), Nagoya, Japan, October 2013.
27. Privacy manipulation and acclimation in a location sharing application. Shomir Wilson, Justin Cranshaw, Norman Sadeh, Alessandro Acquisti, Lorrie Cranor, Jay Springfield, Sae Young Jeong, and Arun Balasubramanian. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* (**UbiComp**), Zurich, Switzerland, September 2013.
28. Tweets are forever: A large-scale quantitative analysis of deleted tweets. Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. In *Proceedings of the 2013 ACM Conference on Computer Supported Cooperative Work* (**CSCW**), San Antonio, TX, February 2013.
29. The creation of a corpus of English metalanguage. Shomir Wilson. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (**ACL**), Jeju, South Korea, July 2012.
30. Application of MCL in a dialog agent. Darsana Josyula, Scott Fults, Michael L. Anderson, Shomir Wilson, and Don Perlis. In *Papers from the Third Language and Technology Conference* (**LTC**), Poznań, Poland, October 2007.

Peer-Reviewed Journal Articles

1. Online Self-Disclosure, Social Support, and User Engagement During the COVID-19 Pandemic. Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. *ACM Transactions on Social Computing* (**TSC**), 2023.
2. Researchers' Experiences in Analyzing Privacy Policies: Challenges and Opportunities. Abraham Mhaidli, Selin Fidan, An Doan, Gina Herakovic, Mukund Srinath, Lee Matheson, Shomir Wilson, and Florian Schaub. In *Proceedings on Privacy Enhancing Technologies* (**PoPETs**) 2023(4), 2023. **Best Student Paper.**
3. Digital Inequality Through the Lens of Self-Disclosure. Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. In *Proceedings on Privacy Enhancing Technologies* (**PoPETs**) 2021(3), 2021.
4. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Sebsatian Zimmeck, Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah Smith. In *ACM Transactions on the Web* (**TWEB**) 13(1), December 2018.
5. PrivOnto: A semantic framework for the analysis of privacy policies. Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. In *Semantic Web Journal* (**SWJ**), May 2017.
6. Nudges for privacy and security: Understanding and assisting users' choices online. Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga

Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. In *ACM Computing Surveys* (**CSUR**) 50(3), August 2017.

7. In search of the use-mention distinction and its impact on language processing tasks. Shomir Wilson. In *The International Journal of Computational Linguistics and Applications* 2(1-2), pp. 139-154, 2011.

Book Chapters

1. Nudges (and Deceptive Patterns) for Privacy: Six Years Later. Alessandro Acquisti, Idris Adjerid, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Yang Wang, and Shomir Wilson. In Sabine Trepte, Philipp Masur (Ed.), *The Routledge Handbook of Privacy and Social Media*, Taylor & Francis, 2023.
2. A bridge from the use-mention distinction to natural language processing. Shomir Wilson. In Saka, P., Johnson, M. (Ed.), *The Semantics and Pragmatics of Quotation*. Springer, 2017.
3. The metacognitive loop and reasoning about anomalies. Matthew Schmill, Michael L. Anderson, Scott Fults, Darsana Josyula, Tim Oates, Donald Perlis, Hamid Haidarian Shahri, Shomir Wilson, and Dean Wright. In Cox, M., Raja, A. (Ed.), *Metareasoning: Thinking About Thinking*. MIT Press, MA, 2010.

Magazine Articles

1. Reports on the 2019 AAAI Spring Symposium Series (Privacy-Enhancing Artificial Intelligence and Language Technologies). Shomir Wilson, et al. *AI Magazine* 40:3, Fall 2019.
2. Reports on the 2016 AAAI Fall Symposium Series (Privacy and Language Technologies). Patrícia Alves-Oliveira, Richard G. Freedman, Dan Grollman, Laura Herlant, Laura Humphrey, Fei Liu, Ross Mead, Frank Stein, Tom Williams, and Shomir Wilson. *AI Magazine* 38:2, Summer 2017.
3. A self-help guide for autonomous systems. Michael L. Anderson, Scott Fults, Darsana P. Josyula, Tim Oates, Don Perlis, Matthew D. Schmill, Shomir Wilson, and Dean Wright. *AI Magazine* 29:2, Summer 2008.

Peer-Reviewed AAAI Symposium Proceedings

1. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. Frederick Liu, Shomir Wilson, Florian Schaub and Norman Sadeh. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*, Arlington, VA, November 2016.
2. Automatic extraction of opt-out choices from privacy policies. Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson and Norman Sadeh. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*, Arlington, VA, November 2016.
3. Analyzing and predicting privacy law compliance of mobile apps. Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin and Joel Reidenberg. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*, Arlington, VA, November 2016.
4. The Metacognitive Loop: An architecture for building robust intelligent systems. Hamid Haidarian, Wikum Dinalankara, Scott Fults, Shomir Wilson, Don Perlis, Matt Schmill, Tim

Oates, Darsana Josyula, and Michael Anderson. In *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge*, Arlington, VA, November 2010.

5. Toward domain-neutral human-level metacognition. Michael L. Anderson, Matt Schmill, Tim Oates, Don Perlis, Darsana Josyula, Dean Wright, and Shomir Wilson. In *Proceedings of the 2007 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Palo Alto, CA, March 2007.

Peer-Reviewed Conference Poster Abstracts

1. Comparing Scam Emails and Email User Education at Universities. Duo Pan, Ellen Poplavska, Nora O'Toole, and Shomir Wilson. In *The Seventeenth Symposium on Usable Privacy and Security (unpublished work)*, held online, August 2021.
2. A Multilingual Comparison of Email Scams. Duo Pan, Ellen Poplavska, Yichen Yu, Susan Strauss, and Shomir Wilson. In *The Sixteenth Symposium on Usable Privacy and Security (unpublished work)*, held online, August 2020.
3. Automatic Title Generation to Improve the Readability of Privacy Policies. Abhijith Athreya Mysore Gopinath, Vinayshekhar Bannihatti Kumar, Shomir Wilson, Norman Sadeh. In *The Sixteenth Symposium on Usable Privacy and Security (unpublished work)*, held online, August 2020.
4. Privacy Not Found: A Study of the Availability of Privacy Policies on the Web. Soundarya Nurani Sundareswara, Shomir Wilson, Mukund Srinath, C. Lee Giles. In *The Sixteenth Symposium on Usable Privacy and Security (unpublished work)*, held online, August 2020.
5. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. Abhilasha Ravichander, Alan Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. In *The Sixteenth Symposium on Usable Privacy and Security (published work)*, held online, August 2020.
6. Increasing the salience of data use opt-outs online. Namita Nisal, Sushain K. Cherivirala, Kanthashree M. Sathyendra, Margaret Hagan, Florian Schaub, Shomir Wilson, Lorrie Faith Cranor, and Norman Sadeh. In *The Thirteenth Symposium on Usable Privacy and Security (unpublished work)*, Santa Clara, CA, June 2017.
7. Mobile app privacy compliance: Automated technology to help regulators, app stores and developers. Sebastian Zimmeck, Lieyong Zou, Bin Liu, Shomir Wilson, Steven M. Bellovin, Ziqi Wang, Roger Iyengar, Florian Schaub, Norman Sadeh, and Joel Reidenberg. In *The Thirteenth Symposium on Usable Privacy and Security (published work)*, Santa Clara, CA, June 2017.
8. Visualization and interactive exploration of data practices in privacy policies. Sushain K. Cherivirala, Florian Schaub, Mads Schaarup Andersen, Shomir Wilson, Norman Sadeh, and Joel R. Reidenberg. In *The Twelfth Symposium on Usable Privacy and Security (unpublished work)*, Denver, CO, 2016.
9. Towards usable privacy policies: Semi-automatically extracting data practices from websites' privacy policies. Norman Sadeh, Alessandro Acquisti, Travis Breaux, Lorrie Cranor, Aleecia McDonald, Joel Reidenberg, Noah Smith, Fei Liu, N. Cameron Russell, Florian Schaub, Shomir Wilson, James Graves, Pedro Leon, Rohan Ramanath, and Ashwini Rao. In *The Tenth Symposium on Usable Privacy and Security (published work)*, Palo Alto, CA, July 2014.

Peer-Reviewed Workshop Proceedings

1. Automated Ableism: An Exploration of Explicit Disability Biases in AIaaS Sentiment and Toxicity Analysis Models. Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. In *Proceedings of the Third Workshop on Trustworthy Natural Language Processing* (TrustNLP), 2023. **Best Short Paper.**
2. Effects of Online Self-Disclosure on Receiving Social Support During the COVID-19 Pandemic. Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. In *1st Workshop on NLP for Positive Impact* (unpublished papers) at ACL, 2021.
3. Demystifying Privacy Policies with Language Technologies: Progress and Challenges. Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy, and Norman Sadeh. In *Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS) at LREC*, Portoroz, Solvenia, May 2016.
4. Distinguishing use and mention in natural language. Shomir Wilson. In *Proceedings of the NAACL HLT Student Research Workshop*, Los Angeles, CA, June 2010.
5. The role of metacognition in robust AI systems. Matt Schmill, Tim Oates, Michael L. Anderson, Darsana Josyula, Don Perlis, Shomir Wilson, and Scott Fults. In *Papers from the Workshop on Metareasoning at the 23rd AAAI Conference on Artificial Intelligence*, Chicago, IL, July 2008.
6. Ontologies for reasoning about failures in AI systems. Michael L. Anderson, Scott Fults, Darsana Josyula, Tim Oates, Don Perlis, Matt Schmill, and Shomir Wilson. In *Proceedings of the First International Workshop on Metareasoning in Agent-Based Systems*, Honolulu, HI, 2007.

Dissertation

- *A Computational Theory of the Use-Mention Distinction in Natural Language.* Shomir Wilson. University of Maryland, 2011.

Reports

1. Case studies and user interaction. Zinaida Benenson, Abdullah Elbi, Zekeriya Erkin, Natasha Fernandes, Simone Fischer-Hübner, Ivan Habernal, Els Kindt, Anna Leschanowsky, Pierre Lison, Christina Lohr, Emily Mower Provost, Jo Pierson, David Stevens, Francisco Teixeira, and Shomir Wilson. In *Privacy in Speech and Language Technology (Dagstuhl Seminar 22342)*, 2023.
2. Towards Automatic Classification of Privacy Policy Text. Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. Technical Report CMU-LTI-17-010, Carnegie Mellon University, 2017.
3. The Usable Privacy Policy Project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Norman Sadeh, Alessandro Acquisti, Travis Breaux, Lorrie Cranor, Aleecia McDonald, Joel Reidenberg, Noah Smith, Fei Liu, N. Cameron Russell, Florian Schaub, and Shomir Wilson. Technical Report CMU-ISR-13-119, Carnegie Mellon University, 2013.
4. Automatic categorization of privacy policies: A pilot study. Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A. Smith. Technical Report CMU-LTI-12-019 / CMU-ISR-12-114, Carnegie Mellon University, 2012.

Other Papers

1. Survey on Sociodemographic Bias in Natural Language Processing. Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J. Passonneau. arXiv:2306.08158, 2023.
2. Effects of Online Self-Disclosure on Social Feedback During the COVID-19 Pandemic. Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. arXiv:2209.10682, 2022.
3. Creation and Analysis of an International Corpus of Privacy Laws. Sonu Gupta, Ellen Poplavska, Nora O'Toole, Siddhant Arora, Thomas Norton, Norman Sadeh, and Shomir Wilson. arXiv:2206.14169, 2022.
4. An Exploratory Analysis of Broadcast Police Communications in Chicago. Pranav Venkit, Chris Graziul, Miranda Goodman, Samantha Kenny, Shomir Wilson. Abstract accepted by the Penn State Annual Social Thought Conference, 2022.
5. Automated Detection of Doxing on Twitter. Younes Karimi, Anna Squicciarini, and Shomir Wilson. arXiv:2202.00879, 2022. Preprint of our CSCW 2022 paper.
6. Identification of Bias Against People with Disabilities in Sentiment Analysis and Toxicity Detection Models. Pranav Narayanan Venkit and Shomir Wilson. arXiv:2111.13259, 2021.
7. A 'Sourceful' Twist: Emoji Prediction Based on Sentiment, Hashtags and Application Source. Pranav Venkit, Zeba Karishma, Chi-Yang Hsu, Rahul Katiki, Kenneth Huang, Shomir Wilson, and Patrick Dudas. arXiv:2103.07833, 2021.
8. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. Mukund Srinath, Shomir Wilson, and C. Lee Giles. arXiv:2004.11131, 2020. Preprint of our ACL 2021 paper.
9. An active logic approach to Moore's Paradox. Shomir Wilson. Scholarly paper for M.S. in Computer Science, 2008.
10. Evaluation of functional-linkage networks applied to protein annotation. Shomir Wilson. Honors thesis for B.S. in Computer Science, 2005.
11. Wittgenstein takes the Turing Test. Shomir Wilson. Selected for presentation at The First Undergraduate Philosophy Conference at Northwestern University, 2005.
12. The pathological liar: An exclusionary approach to self-referential contradictions in natural language. Shomir Wilson. *Aporia* 14(2), 2004.
13. Construction of a crystal graph simulation engine. Shomir Wilson. Honors thesis for B.S. in Mathematics, 2004.

PUBLICLY AVAILABLE DATASETS AND TOOLS

University Scam Emails Corpus: 5,155 scam messages sent to US university email addresses. [hyperlink]. Documented with collaborators in my WebConf 2023 publication.

GPI (Government Privacy Instructions) Corpus: 1,043 privacy laws, regulations, and guidelines from 182 jurisdictions around the world. [hyperlink]. Documented with collaborators in my ArXiv 2022 publication.

PrivaSeer Corpus: 1M web privacy policies, covering a larger number of websites than any other privacy policy corpus to date. [hyperlink]. Documented with collaborators in my ACL 2021 publication.

PrivacyQA Dataset: The first dataset to support automated question answering in the privacy policy domain [hyperlink]. Documented with collaborators in my EMNLP 2019 publication.

PrivaSeer: The first privacy policy search engine, enabling privacy policy research at web scale [hyperlink]. Documented with collaborators in my 2021 ICWE paper.

Opt-Out Easy: A web browser plugin to help internet users take control of their privacy [hyperlink]. Documented with collaborators in my Web Conference 2020 publication.

ASDUS (Automatic Segment Detection using Unsupervised and Supervised Learning): A tool for extracting section titles and prose text from HTML independently of underlying differences in markup [hyperlink]. Documented with collaborators in my EMNLP 2018 publication.

Usable Privacy Policy Explore Site: Showcasing interpretations of privacy policies by humans and by machine learning [hyperlink]. Documented with collaborators in my ACL 2016 publication.

OPP-115 (Online Privacy Policies, Set of 115) Corpus: The first corpus of privacy policies to feature annotations by legal experts. [hyperlink]. Documented with collaborators in my ACL 2016 publication. **October 2021: Commercial licenses purchased by a Big Five information technology company and a major cybersecurity company.**

Enhanced Cues Corpus: The first corpus of English metalanguage [hyperlink]. Documented in my ACL 2012 publication.

INVITED ACTIVITIES

Invited Talks

"Sociodemographic Biases in Natural Language Processing: Two Case Studies", Language Meets Technology Speaker Series, Center for Language Science, Pennsylvania State University, 10 February 2023.

"Natural Language Processing for Privacy and Social Good", Privacy Seminar, Carnegie Mellon University, 31 January 2023.

"Natural Language Processing for Privacy and Beyond", Privacy Engineering - Regulatory Compliance Lab, University of Maine, 21 April 2022.

"Human Language Technologies for Understanding Online Privacy", Computer Science Seminar, Dalhousie University, 5 December 2018.

"Human Language Technologies for Understanding Online Privacy", Data Sciences Seminar, Penn State College of Information Sciences and Technology, 27 November 2018.

"Natural Language Processing for the Privacy of Internet Users", Northern Kentucky University College of Informatics, 24 March 2017.

"Text Analysis to Support the Privacy of Internet Users", Hutton Lecture Series, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 17 February 2017.

"Natural Language Processing to Support Internet Privacy", University of Dayton Computer Science Department Colloquium Series, 3 February 2017.

"Text Analysis to Support the Privacy of Internet Users", Georgetown University Computer

Science Colloquium Series. 21 November 2016.

"Crowdsourcing Annotations of Websites' Privacy Policies: Can It Really Work?", **Encore Track** at the Fourth AAAI Conference on Human Computation and Crowdsourcing, Austin, Texas. 1 November 2016.

"Introspective Users and Introspective Text: Some Recent Results", CHIME Text Seminar, National University of Singapore. 5 January 2016.

"Identifying Deixis to Communicative Artifacts in Text", NLIP Seminar Series, Cambridge University. 9 May 2014.

"An Empirical Approach to Metalanguage", ILCC/HCRC Seminar Series, University of Edinburgh. 6 September 2013.

"A Computational Approach to Metalanguage and the Use-Mention Distinction", CL+NLP Lunch, Carnegie Mellon University. Pittsburgh, PA, 23 April 2013.

"Distinguishing Use and Mention in Natural Language", Centre for Language Technology Seminar Series, Macquarie University. Sydney, Australia, 2009.

Other Invited Activities

Participant in CRA LEVEL UP Workshop (event to gather best practices for inclusive undergraduate computing education), 2023-08-(07-08).

Participant in NSF SaTC Vision 2.0 Workshop, 2023-03-(08-09).

Panelist in "Tenure-Track Interview Panel Discussion for Postdocs", organized by Penn State's Office of the Senior Vice President for Research, 2023-01-20.

Participant in Dagstuhl Seminar "Privacy in Speech and Language Technology", 2022 August 21-26.

Panelist for Forums on "Meta-Level Control" and "Models of Self", during the Metareasoning Workshop at the Twenty-Third AAAI Conference on Artificial Intelligence, 2008.

MISCELLANY

Professional Memberships: AAAI, ACL, ACM.

AED/CPR training, 2023, 2021, 2019.

QPR training (for suicide intervention), 2020.

Safer People, Safer Places training (for LGBT+ inclusiveness), 2018.

Opiod Overdose Prevention training, 2017.

Bias Busters training (for implicit bias identification and intervention), 2015.

Licensed amateur radio operator, technician class.