Shahriar Shayesteh

PhD Student, Pennsylvania State University, USA
shahriar.shayesteh92@gmail.com

RESEARCH FOCUS

I study how control over personal data shifts as the web moves from static, policy-governed websites to an agentic web where AI agents act on users' behalf. My research analyzes how privacy policy disclosure norms differ across industries and how these norms evolve as AI agents replace direct user–website interaction, using large-scale sectoral analysis of websites, privacy policies, and agent tool use.

SELECTED PUBLICATIONS

Generative Adversarial Learning with Negative Data Augmentation for Semi-Supervised Text Classification.
Shahriar Shayesteh, Diana Inkpen. *FLAIRS-35, 2022.*

SoACer: Sector-Based Corpus & LLM Framework for Sectoral Website Classification.
Shahriar Shayesteh, Mukund Srinath, Lee Matheson, Lu Xian, Sinjoy Saha, Lee Giles, Shomir Wilson. *ACM DocEng, 2025.*
(Enables large-scale empirical analysis of sectoral data practices for governance and oversight research.)

The PrivaSeer Project: Large-Scale Resources for Analysis of Privacy Policy Text.
Shomir Wilson, Florian Schaub, Lee Matheson, Shahriar Shayesteh, Lu Xian. *USENIX SOUPS, 2025.*
(Provides infrastructure for studying compliance, transparency, and data governance at scale.)

Privacy, More or Less: Large-Scale Sectoral Comparison of Privacy Policies Between Industries *(Submitted to LREC 2026)*
Constructed a longitudinal corpus of 59K privacy policies to analyze sector-level norms, deviations, and mismatches between stated policies and observed practices.

Research and Internship Experience

Graduate Research Assistant Fall 2023 – Present
The Human Language Technologies Lab, Pennsylvania State University

- **Project: PrivaSeer** — a large-scale platform for collecting, indexing, and analyzing privacy policy text.
- **Action:** Designed and implemented sector-aware analysis over 3M privacy policies, including extraction of data types, purposes, and third-party sharing practices.
- **Outcome:** Enabled cross-sector comparison of privacy practices and supported empirical research on how organizations state and vary data handling commitments.

Research Intern Feb 2022 – Apr 2022
Department of Canadian Heritage

- **Problem:** Understanding systemic challenges faced by artists using unstructured qualitative data.
- **Action:** Applied NLP methods (topic modeling, sentiment analysis) to analyze large text corpora; translated findings into policy-relevant insights.
- **Result:** Informed cultural policy design and strategic planning for government stakeholders.

Graduate Research Assistant Jan 2021 – Jun 2023
NLP Laboratory, University of Ottawa

- **Problem:** Social bias and reliability concerns in semi-supervised text classification.
- **Action:** Investigated fairness-aware training methods; implemented a GAN-based approach with negative data augmentation.
- **Result:** Achieved +3% performance improvement over strong baselines; work formed the basis of a Master's thesis.

RAI Summer School at Mila Jun 2023
Mila Institute – Responsible AI and Human Rights

- **Focus:** Ethical, legal, and governance challenges in AI systems.
- **Outcome:** Engaged in interdisciplinary discussions on accountability, fairness, and human-centric AI design.

Governance & Policy Engagement

- Experience translating technical AI risks (agent behavior, tool misuse, data flows) into governance-relevant insights.
- Familiarity with privacy regulation contexts (e.g., consent, purpose limitation) through large-scale policy text analysis.
- Research aligned with AI accountability, transparency, and responsible deployment.

Education

Pennsylvania State University Fall 2023 – Fall 2027 (Expected)
Ph.D. in Informatics GPA: 3.93/4.0

University of Ottawa Jan 2021 – Jun 2023
M.Sc. in Computer Science GPA: 4.0/4.0

University of British Columbia Sep 2017 – May 2020
Non-Degree Program, Computer Science and Statistics GPA: 3.73/4.0

Ongoing Research Projects

TOOLS-R: Robustness Diagnostics for Tool-Calling LLMs. *(Ongoing)*
Studying how LLM-based agents fail when selecting and invoking tools, with a focus on reliability, misuse, and error propagation in real deployments.

LLM-COLDs: Consumer-Oriented Legal Document Summarization. *(Ongoing)*
Exploring LLM-based and agentic methods to generate user-centric summaries from consumer-based legal documents.

References

Prof. Shomir Wilson
Associate Professor, IST, Pennsylvania State University
Email: shomirwilson@psu.edu