Research Statement

R. Teal Witter

I am a theoretical computer scientist studying algorithms with an eye towards how they can positively impact society. As computing becomes ubiquitous, the design and analysis of algorithms for social good is increasingly important. We want to ensure that algorithms are applied in useful ways (e.g., improving healthcare, education, and transportation) while simultaneously ensuring that algorithms are unbiased (e.g., decisions in hiring, credit loans, and criminal justice are fair and transparent). The growing importance of this area is reflected in the development of recent venues, including the FAccT conference and social impact tracks at major machine learning conferences like AAAI and IJCAI.

I have studied algorithms for social good in the context of explainable AI, evaluation of nonprofit efficacy, fairness in machine learning, resource allocation, and societal polarization. I leverage a broad theoretical toolkit including techniques in randomized linear algebra, linear programming, and the theory of boolean functions. But work in my area also requires deep interdisciplinary engagement with practitioners and stakeholders. To this end, I have worked closely with an early childhood literacy nonprofit and collaborated with researchers across nine institutions, publishing in top venues like NeurIPS, AAAI, and ESA.

Algorithms for Explainable AI

As AI predictions are increasingly incorporated into high-stakes domains, users and auditors of AI systems should understand why a prediction was made. For example, a credit card applicant should know why their application was rejected, and a defendant should be aware of how their bail was set. In broader societal applications, such as government spending or nonprofit resource allocation, explainability becomes even more critical. It's not enough to explain individual predictions; stakeholders should have confidence in the entire model's transparency and reasoning. For example, an early childhood literacy nonprofit benefits from a transparent, simple model to evaluate the impact of their program, allowing them to trust the analysis and use it to guide future decisions

In recent work, I empirically and theoretically improved one of the most popular methods for explaining AI predictions. Shapley values are one of the primary methods in explainable AI, quantifying how changing input features affects model output. In recent work, I used a theoretically motivated technique called leverage score sampling to both empirically and theoretically improve the state-of-the-art Kernel SHAP estimator [MW24]. The algorithm I proposed, Leverage SHAP, gives better empirical performance than even the highly optimized official implementation and offers theoretical guarantees, contrasting with Kernel SHAP. In follow-up work, I applied the same leverage score sampling technique to a related but more robust game-theoretic called Banzhaf values [LWK+24]. Together, my work establishes more efficient and theoretically motivated methods for explaining AI predictions.

A major motivation of computing Shapley values is to add transparency to predictions, so we can either detect unfair decision-making or verify that methods are fair. Fairness is an important topic in machine learning and a major technical and philosophical question is how fairness should be defined and measured. I have contributed research on how to define notions of fairness [RW23] and measure fairness in the presence of unavoidable uncertainty [RW24].

A key tenet of explainable AI is algorithmic simplicity, which ensures models are both interpretable and reliable. In collaboration with the early childhood literacy nonprofit Reach Out and Read Colorado (RORCO), I have applied this principle to the challenge of treatment effect estimation. While treatment effect estimation is well-studied, existing algorithms are often complex and yield inconsistent estimates. To address this, I developed a benchmark for evaluating treatment effect estimators and proposed a theoretically-motivated, simple method [WM24]. By leveraging regression tools related to my work on Shapley and Banzhaf values, I introduced a simple yet accurate algorithm that RORCO has already deployed.

Online Decision Making in a Dynamic World

My work on explainable AI focuses on algorithms that make predictions based on static data. However, many of the most compelling applications of algorithms for social good involve dynamic systems, where models repeatedly interact with the world. For example, a traffic optimization algorithm suggests a street to open, observes the resulting vehicle flow, and then adapts its next suggestion based on the new conditions. Similarly, an algorithm for reintroducing endangered species makes habitat recommendations, monitors the species' success, and refines its future suggestions. A major second thread of my work focuses on designing algorithms for these dynamic problems.

The NYC Open Streets Project (closing streets to cars, opening streets to people) is a cost-effective method to modify urban infrastructure. To optimize which streets are opened, I designed a deep reinforcement learning model that incorporates both temporal and spatial data, allowing it to adapt to the city's complex traffic environment while balancing the dual objectives of minimizing congestion and reducing collisions [WR24]. By integrating multiple data sources—such as traffic patterns, accident reports, and weather—the model optimizes with a granular view of urban mobility. Developed in collaboration with infrastructure experts, this approach serves as a proof-of-concept for solving dynamic, socially impactful problems, with potential for broad application in urban planning.

Many dynamic problems, such as reintroducing endangered species, involve achieving specific goals while minimizing the cost of actions. These settings can be modeled as resource allocation problems to restless bandits, where actions impact constantly changing environments. The standard formulations of restless multi-armed bandits typically focus on maximizing impact within a cost budget, but they overlook scenarios like wildlife conservation, where achieving a positive impact is the primary constraint. I propose a dual formulation of the restless bandits problem that prioritizes achieving the goal while minimizing costs and show that solving it—even approximately—is PSPACE-hard [WH24]. My work lays the foundation for approaching dynamic restless bandit problems with goal constraints, highlighting the need for novel algorithms. Additionally, I have fundamental algorithmic work on other resource allocation problems, such as optimizing the order of actions to maximize reward within a cost framework. In this context, I analyzed an evolutionary algorithm for the Min-Sum Submodular Cover Problem, demonstrating that it provides similar theoretical guarantees as the standard greedy algorithm while offering more diverse solutions [HLW22].

Sometimes our goal is not to develop an algorithm, but to use algorithmic tools to model a dynamic process that we observe in the real world. For example, we qualitatively observe that politics is becoming more polarized but we do not have a simple opinion dynamics model for understanding this phenomenon. Prior works have developed increasingly complicated models that exhibit polarization with different contrived dynamics. I took a theoretical lens to view one of the simplest opinion dynamics models under a scale-invariant measure of polarization. In this view, the simple opinion dynamics model exhibits relative polarization, reflecting the phenomenon of political polarization [MRUW22].

Future Work

While my main research has centered on algorithms for social good, I remain curious about new topics in theoretical computer science and consider myself a generalist. For instance, I have worked on algorithms for efficiently evaluating Boolean functions in both classical [HKLW22] and quantum settings [CKW23, KW21, DKW19]. Looking ahead, I hope to continue exploring these and other areas at ETH Zurich ITS, since the topics offer excellent opportunities for undergraduate research. My own journey into computer science research began with strategies for the board game Ticket-to-Ride [WL20] and the computational complexity of Backgammon [Wit21], illustrating how accessible topics can spark interest in the field.

By design, my research agenda is multi-faceted, combining theoretical analysis and motivation of algorithms with a focus on practical efficiency and real-world impact. This approach enables students to carve out projects that align with their interests and strengths. Students with a strong mathematical background can leverage creative ideas to design novel algorithms, refining them for theoretical analysis. Students with strong computational skills can focus on efficiently implementing algorithms, identifying and addressing practical concerns. I have advised four undergraduate and high school students on research projects, and I'm excited to continue involving students in my future work across the following research agenda.

Explainable AI Estimators Beyond Shapley and Banzhaf Values

Interpreting predictions is key to building transparency and trust in AI systems. In recent work, I designed theoretically-motivated algorithms for estimating game-theoretic explanations of AI predictions, specifically Shapley and Banzhaf values. To enhance these methods, I applied leverage score sampling, exploiting elegant connections between the respective quantities and linear regression problems. There are many game-theoretic quantities beyond Shapley and Banzhaf values that each satisfy different properties that may be particularly relevant to different explainable AI tasks. I plan to identify quantities relevant to different explainable AI tasks and search for connections to linear regression, ideally employing leverage score sampling to the special structure of the problem and yielding efficient algorithms.

Active Sampling for Treatment Effect Estimation

Simple and accurate treatment effect estimation is key to the decision making of charitable and nonprofit organizations. My treatment effect estimation work with Reach Out and Read Colorado (RORCO) focused on the natural experiment setting where treatments have already been applied. In this setting, I used regression adjustment to design better algorithms. I plan to go beyond natural experiments to the active setting where individuals are selected to either receive the 'treatment' or 'control'. I hope to apply leverage score sampling to design efficient estimators that can achieve the same level of accuracy but with fewer selected individuals, limiting the negative impact to individuals that are excluded from the nonprofit's 'treatment'.

Distortion-free Watermarking for Responsible AI

AI is increasingly relevant around us. Large Language Models (LLMs) have become ubiquitous for text generation and even high-quality images can now be generated from AI models. Responsibly using these tools requires differentiating between human and AI generated content, e.g., to prevent plagiarism or detect malicious actors. Standard techniques for "watermarking" content modify the outputs, potentially allowing malicious actors to fake watermarks. I am interested in distortion-free approaches where the content is generated from the true distribution, and it is only possible to verify the watermark with access to a private correlated variable [AFW+24]. The downside of current distortion-free methods is that the correlated random variable must be stored. I plan to leverage information already present in the image or text to robustly and securely store the correlated variable, enabling efficient distortion-free watermarking and ultimately supporting the responsible use of AI.

Conclusion

Algorithms are all around us, making our lives better but sometimes introducing biases and harm. My work seeks to improve transparency, explaining the way these models work , designing simple yet effective algorithms that can be trusted by stakeholders, and adding guard rails against the misuse of AI. I leverage both mathematical tools and algorithmic insights to solve impactful problems, iteratively identifying and solving problems with stakeholder input. I am particularly excited to further incorporate students in my research, carving out impactful problems that strengthen the skills of student researchers and encourage learning.

References

An asterisk (*) indicates that authors are listed in alphabetical order.

[AFW+24] Kasra Arabi, Benjamin Feuer, R Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. In *Submission*, 2024.

[CKW23] Michael Czekanski*, Shelby Kimmel*, and R Teal Witter*. Robust and space-efficient dual adversary quantum query algorithms. In *European Symposium on Algorithms*, 2023.

[DKW19] Kai DeLorenzo*, Shelby Kimmel*, and R Teal Witter*. Applications of the quantum algorithm for st-connectivity. In *Conference on the Theory of Quantum Computation, Communication and Cryptography*, 2019.

[HKLW22] Lisa Hellerstein*, Devorah Kletenik*, Naifeng Liu*, and R Teal Witter*. Adaptivity gaps for the stochastic boolean function evaluation problem. In *Workshop on Approximation and Online Algorithms*, 2022.

[HLW22] Lisa Hellerstein*, Thomas Lidbetter*, and R Teal Witter*. A local search algorithm for the min-sum submodular cover problem. In *International Symposium on Algorithms and Computation*, 2022.

[KW21] Shelby Kimmel* and R Teal Witter*. A query-efficient quantum algorithm for maximum matching on general graphs. In *Algorithms and Data Structures Symposium*, pages 543–555, 2021.

[LWK+24] Yurong Liu*, R Teal Witter*, Flip Korn, Tarfah Alrashed, Dimitris Paparas, and Juliana Freire. Kernel banzhaf: A fast and robust estimator for banzhaf values. In *Submission*, 2024.

[MRUW22] Christopher Musco*, Indu Ramesh*, Johan Ugander*, and R Teal Witter*. How to quantify polarization in models of opinion dynamics. In *International Workshop on Mining and Learning with Graphs*, 2022.

[MW24] Christopher Musco* and R Teal Witter*. Provably accurate shapley value estimation via leverage score sampling. In *Submission*, 2024.

[RW23] Lucas Rosenblatt and R Teal Witter. Counterfactual fairness is basically demographic parity. In *AAAI Conference on Artificial Intelligence*, 2023.

[RW24] Lucas Rosenblatt* and R. Teal Witter*. Fairlyuncertain: A comprehensive benchmark of uncertainty in algorithmic fairness. In *Submission*, 2024.

[WH24] R Teal Witter and Lisa Hellerstein. Minimizing cost rather than maximizing reward in restless multi-armed bandits. In *Submission*, 2024.

[Wit21] R Teal Witter. Backgammon is hard. In *International Conference on Combinatorial Optimization and Applications*, 2021.

[WL20] R Teal Witter and Alex Lyford. Applications of graph theory and probability in the board game ticket to ride. In *International Conference on the Foundations of Digital Games*, 2020.

[WM24] R Teal Witter and Christopher Musco. Benchmarking estimators for natural experiments: A novel dataset and a doubly robust algorithm. In *Conference on Neural Information Processing Systems*, 2024.

[WR24] R Teal Witter and Lucas Rosenblatt. I open at the close: A deep reinforcement learning evaluation of open streets initiatives. In *AAAI Conference on Artificial Intelligence*, 2024.