Research Statement

R. Teal Witter

The rapid integration of computing into society has led to algorithms shaping crucial decisions in areas like healthcare, education, and criminal justice. While these algorithmic solutions promise unprecedented societal advancements, they also carry the risk of amplifying biases and creating unintended negative impacts on vulnerable populations. My research focuses on the design and analysis of trustworthy algorithms. To this end, my work follows a systematic progression towards algorithmic systems that are:

1. **Explainable**: I develop efficient algorithms to interpret complex models, enabling stakeholders to understand the rationale behind algorithmic decisions.
2. **Fair**: I design methods to measure algorithmic fairness and mitigate biases, ensuring equitable treatment across diverse populations.
3. **Effective**: I build robust algorithms for socially impactful problems, from resource allocation to nonprofit evaluation, with provable performance guarantees.
4. **Secure**: I formulate safeguards against the misuse and manipulation of AI, including watermarking methods to ensure content authenticity and traceability.

My work draws on ideas from theoretical science to solve problems with social impact, providing (A) **rigorous guarantees on algorithmic performance and behavior**, and (B) **theoretical insights for the design of trustworthy algorithms**. By bridging the gap between theory and practice, I enhance the reliability and impact of algorithms for the social good.

In my work, I leverage a broad theoretical toolkit including techniques in randomized linear algebra, linear programming, and discrete optimization. Research in my area also requires interdisciplinary engagement with stakeholders. To this end, I have worked closely with an early childhood literacy nonprofit and collaborated with researchers across nine institutions, publishing in top machine learning and theoretical computer science venues such as NeurIPS, ESA, and AAAI. Looking ahead, I aim to address emerging challenges in machine learning, particularly focusing on explainable AI for unstructured models and distortion-free watermarking for the responsible use of AI.

Algorithms for Trustworthy Machine Learning

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.

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 of Kernel SHAP, and offers theoretical guarantees. In follow-up work, I applied the same leverage score sampling technique to a related but more robust game-theoretic quantity 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 to strategically allocate resources and improve program efficacy.

Online Decision Making with Social Impact

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 proposed a dual formulation of the restless bandits problem that prioritizes achieving a goal while minimizing costs [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]. This work provides a theoretical foundation for understanding the dynamics of polarization and suggests that simple models can capture complex real-world phenomena.

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]. My journey into computer science research began with strategies for the board game Ticket-to-Ride [WL20] and the computational complexity of Backgammon [Wit21].

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 student collaborators to carve out projects that align with their interests and strengths. 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

My prior work establishes more efficient and theoretically motivated methods for explaining AI predictions with Shapley and Banzhaf values. Building on this foundation, I plan to explore a broader spectrum of game-theoretic concepts applicable to various social good domains. My goal is to leverage my theoretical expertise to design novel, computationally efficient algorithms for these game-theoretic quantities, thereby enhancing the interpretability and transparency of AI systems across diverse applications.

One under-studied social good setting is graph learning tasks, such as predicting the spread of disease or identifying collusion rings. Graph neural networks have emerged as a powerful tool for learning on graph data, processing the features of adjacent nodes to identify local patterns. Because of the high stakes of social good applications, it is important to explain how graph neural networks make predictions. Unfortunately, standard game-theoretic attribution quantities like Shapley and Banzhaf values do not take into account the graph structure, and so lose the ability to explain how the graph neural network reasons. An alternative game-theoretic quantity designed specifically for graph structures is the Hamiache-Navarro (HN) value, which naturally generalizes Shapley values to graph settings. Mathematically, the HN value is the limit of a series of associated games, which can be represented as repeated matrix multiplication. While the HN value satisfies many desirable properties, it is computationally expensive to compute. I plan to investigate the structure of the HN value to analytically derive a more efficient algorithm to compute it, leveraging insights from my prior theoretical work.

There is a rich game theory literature to describe attribution techniques from an axiomatic perspective. As trustworthy AI becomes increasingly important, this literature is a powerful resource for explaining AI predictions in an axiomatic way. Unfortunately, the majority of prior work that adapts game-theoretic quantities to AI applications develops heuristic algorithms. However, as evidenced by my work on Shapley and

Banzhaf values, theoretically motivated algorithms can outperform heuristic methods, while simultaneously offering strong non-asymptotic guarantees. I plan to apply my theoretical toolkit to design provably efficient algorithms for computing game-theoretic quantities relevant to social good applications.

Distortion-free Watermarking for Responsible AI

As AI models become more advanced, powerful tools like Large Language Models (LLMs) for text generation and diffusion models for prompt-guided image generation are ubiquitous. While these technologies have numerous applications, they also bring new risks, such as malicious actors claiming AI-generated text as their own or fabricating realistic images of fake events, causing confusion or even harm. To mitigate these risks, model owners use watermarking techniques to track the content generated by their models. However, most current watermarking methods are distortion-based, meaning they modify the output to embed identifiable markers. Despite their widespread use, distortion-based watermarks remain vulnerable: they can be detected and even forged by malicious actors if enough examples are available.

I plan to develop distortion-free watermarks that scale without the need for additional storage. The key idea is to leverage context to robustly and securely generate a seed using locality-sensitive hashing. This would allow us to generate a correlated random variable from the seed in a secure and efficient manner, without distorting the distribution of the generated content.

In the LLM setting, text is generated in an auto-regressive way by predicting the next token based on the previous context. Current watermarking approaches modify the next-token distribution, either globally or contextually, making the watermarks detectable, removable, and sometimes even noticeably degrading text quality. I plan to use SimHash to convert the embedding of the prompt into seeds and then generate a random variable using cryptographic hash functions. This random variable is reproducible if we have access to the seed, and it is distributed identically to a sample from the true probability distribution over next tokens. To detect the watermark, we compute the correlated random variable for each seed and check its alignment with the generated text. By using SimHash, we can guarantee that nearby embedded contexts produce the same seed with high probability, making the watermark detectable without degrading text quality. A similar approach can be applied to image watermarking. Instead of sampling the next token, we sample random noise to build the image via diffusion.

Combining distortion-free and searchable watermarking with streaming and randomized algorithms promises security and efficiency. I plan to leverage information already present in the image or text to robustly store the correlated variable in the generated content, 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 address 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

In the tradition of theoretical computer science, an asterisk (*) indicates that authors are listed in alphabetical order.

[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.