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 aims to bridge the gap between theoretical foundations and practical applications, ensuring that algorithms designed for social good are not only effective but also trustworthy and explainable. To this end, my work focuses on the following interconnected goals:

1. **Explainable AI**: Developing rigorous mathematical frameworks to interpret complex models, enabling users to understand the rationale behind algorithmic decisions.
2. **Responsible Use of AI**: Formulating safeguards to prevent AI misuse and ensure accountability. This includes watermarking AI-generated content to ensure authenticity and traceability.
3. **Effective Algorithms**: Designing and analyzing algorithms that address societal challenges while maintaining provable guarantees on fairness, efficiency, and robustness. A key social good application is creating tools for nonprofits to rigorously evaluate and optimize the impact of their initiatives.

The growing importance of algorithms for social good 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. Many algorithms proposed for social good in these venues and others are heuristic in nature. My work aims to illuminate their theoretical foundations, providing (A) **rigorous guarantees on algorithmic performance and behavior**, and (B) **theoretical insights for the design of more effective and trustworthy algorithms**. By bridging the gap between theory and practice, I strive to enhance the reliability and impact of algorithms designed for social good.

I have studied algorithms for social good in the context of explainable AI [MW24, LWK+24], evaluation of nonprofit efficacy [WM24], fairness in machine learning [RW23, RW24], resource allocation [HLW22, WR24, WH24], and societal polarization [MRUW22]. 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 practitioners and stakeholders. To this end, I have worked closely with a 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.

In the remainder of this document, I outline my research plans in relation to explainable AI, the responsible use of AI, and effective algorithms for social good. Informed by my prior work, I discuss concrete ideas for future research and opportunities for collaboration at the Sante Fe Institute.

Provably Accurate Algorithms for Explainable AI

As AI predictions are increasingly incorporated into high-stakes domains like finance, law, and healthcare, 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. Further, explaining AI predictions can help identify biases and the patterns learned by models, supporting the improvement and refinement of future algorithms.

My prior work establishes more efficient and theoretically motivated methods for explaining AI predictions with Shapley and Banzhaf values [MW24, LWK+24]. 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.

Research Interests

R. Teal Witter

Distortion-free Watermarking for the Responsible Use of AI

As AI models become more advanced, tools like Large Language Models (LLMs) for text generation and diffusion models for prompt-guided image generation surround us. While these technologies have numerous applications, they also bring new risks, e.g., malicious actors claiming AI-generated text as their own or fabricating realistic images of fake events, potentially causing confusion or harm. To mitigate these risks, model owners use watermarking 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.

Building on my vision watermarking work [CKW23], I plan to integrate distortion-free watermarking and locality sensitive hashing to enhance both security and efficiency. By utilizing existing information within images or text, I aim to embed correlated variables robustly into the generated content. This method enables effective distortion-free watermarking, supporting responsible AI usage while maintaining efficiency.

Simple and Trustworthy Algorithms for Treatment Effect Estimation

In broader societal applications, such as government spending or nonprofit resource allocation, interpretability is even more critical. It's not enough to explain individual predictions; stakeholders should have confidence in the entire model's transparency and reasoning. This need for transparency extends to the realm of treatment effect estimation, an important problem in evaluating the impact of social programs. While randomized control trials often allow us to estimate the effect of a treatment, they're not always possible or ethical to implement. In some cases, certain individuals may have a greater need for the treatment, or the treatment may have already been assigned, leaving us only to observe the outcome. These scenarios, known as natural experiments, are common in social good applications and require sophisticated analytical approaches because the treatment assignment can be confounded by other factors.

While treatment effect estimation is well-studied, there are practically no algorithms with non-asymptotic, user-friendly guarantees. My goal is to develop simple algorithms with understandable theoretical guarantees for treatment effect estimation, building on my prior work with an early childhood literacy nonprofit [WM24]. Like the guarantees I developed for Shapley and Banzhaf estimators, the guarantees would be of the form: with $\text{poly}(m, 1/\epsilon, 1/\delta)$ samples, we can guarantee ϵ-approximate estimates with probability $1 - \delta$. This approach not only enhances the reliability of the estimates but also allows stakeholders to easily evaluate the impact of their programs.

Conclusion

While my primary focus is on algorithms for social good, I approach this field with the breadth of a generalist. This diverse background allows me to bring novel perspectives to socially impactful problems. For instance, my work on efficient Boolean function evaluation in classical [HKLW22] and quantum settings [CKW23, KW21, DKW19] has honed my skills in algorithmic optimization. Even my explorations into the computational complexity of board games [WL20, Wit21] have yielded insights applicable to real-world decision-making processes. These varied experiences continually enrich my approach to core problems in algorithms for social good.

The increasing prevalence of AI and algorithmic decision-making in crucial societal domains underscores the need for methods that are both effective and trustworthy. My research aims to bridge the gap between theoretical computer science and practical, socially relevant applications. I focus on developing algorithms that are not only efficient but also explainable, secure, and interpretable. Through my work on explainable AI, watermarking, and treatment effect estimation, I address some of the most pressing challenges posed by modern algorithmic systems. At the Sante Fe Institute, I look forward to collaborating with fellow researchers to expand the reach of these ideas and create tangible benefits for society.

References

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.