Research Agenda

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 need algorithms that are transparent (e.g., a credit card applicant knows why their application was rejected) and effective (e.g., a model accurately predicts the impact of a nonprofit program), with adequate guard rails (e.g., AI-generated can be tracked back to its source). 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. 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 such as NeurIPS, AAAI, and ESA.

Explaining AI Predictions

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 recent work, I empirically and theoretically improved one of the most popular methods for explaining AI predictions. Shapley values quantify how changing input features affects model output. (The SHAP paper has more than 25,000 citations and the associated codebase has been used in almost 20,000 Github projects.) One of the most popular and efficient model-agnostic methods for computing Shapley values, Kernel SHAP, exploits an elegant mathematical connection to linear regression but in a heuristic way. In recent work, I used a theoretically motivated technique called leverage score sampling to both empirically and theoretically improve Kernel SHAP [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 approach called Banzhaf values [LWK+24]. Together, my work establishes more efficient and theoretically motivated methods for explaining AI predictions.

I will take theoretical techniques to other game-theoretic quantities relevant in social good applications. For example, Shapley values sum to the model output, making them desirable for explaining individual predictions. Meanwhile, Banzhaf values satisfy an efficiency property that is useful when features are composed of many sub-features (e.g., a loan applicant's net worth is the sum of their assets and liabilities). I will build collaborations with stakeholders to identify the settings where they would benefit from transparent and trustworthy AI predictions. Once I identify the properties most relevant for the stakeholder, I will investigate the structure of the problem and apply tools from my extensive theoretical toolkit to design efficient algorithms. In this way, I plan to combine interdisciplinary engagement with theoretical analysis to solve impactful explainable AI problems.

Effective Algorithms for Social Applications

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. 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 used to inform future program development.

While my work uses existing data to approximate the impact of a program, the quality of the predictions is inherently limited by the confounded nature of the data. In future work, I plan to extend my treatment effect estimation work to the active regression setting, where individuals are selected to either receive the 'treatment' or 'control'. I will apply theoretical tools to design efficient estimators that can achieve the same level of accuracy as prior work 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

As data processing and computing techniques quickly advance, AI models are increasingly prevalent. Large Language Models (LLMs) have become ubiquitous for text generation and AI models can generate realistic images from text prompts. While their applications are vast, these new capabilities also introduce new challenges: Malicious actors may claim that AI generated text is their own or, worse, use AI models to generate images of fake events, causing confusion or even harm. Currently, model owners like OpenAI or Google use watermarking techniques to track the source of generated content. However, the current watermarking techniques are distortion-based, meaning they modify the output to embed the watermark which, in turn, can be detected and forged. I am interested in distortion-free watermarking techniques where the content is generated without modification. In this setting, it is only possible to verify the watermark with access to a private correlated variable [AFW+24]; hence, the watermark is secure and robust to forgery. The downside of current distortion-free methods is that the correlated variable must somehow be stored by the model owner, which can be costly and inefficient. I plan to leverage information already present in the image or text to robustly and securely store the correlated variable in the generated content, enabling efficient distortion-free watermarking and ultimately supporting the responsible use of AI.

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. I have advised four undergraduate and high school students on research projects, and I'm excited to continue involving students in my research at Bowdoin College through initiatives like the Freedman and Weinberger Research Fellowships.

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

**As is the custom in theoretical computer science, authors are listed in alphabetical order unless otherwise specified with an asterisk.*

[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.

[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.

[MW24] Christopher Musco and R Teal Witter. Leverage shap: Estimating shapley values with leverage score sampling. In *Submission*, 2024.

[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.