# 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 want to ensure that algorithms are transparent (e.g., credit card applicant should know why their application was rejected), effective (e.g., the model should accurately predict the impact of a nonprofit program), and used responsibly (e.g., the content generated by AI 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 like NeurIPS, AAAI, and ESA.

My current research is divided into two main threads: algorithms for explainable AI and effective algorithms for social applications. In the former, I focus on developing algorithms that provide transparency and trust in AI predictions. In the latter, I design state-of-the-art algorithms for some of the most compelling applications for social good. I have also worked on a variety of other topics in theoretical computer science, including quantum computing, evolutionary algorithms, and opinion dynamics. In addition to my current work, the third prong of my research agenda is the responsible use of AI. As AI develops at a break-neck speed, there are few guardrails on the use of image and text generation tools. I plan to develop theoretically-motivated algorithms to identify AI generated content, ensuring that AI is used responsibly and ethically.

## 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 are one of the primary methods in explainable AI, quantifying 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.

Shapley value satisfy four properties. Banzhaf values a set of four properties. Properties important for different applications. I plan to build collaborations with stakeholders and nonprofits, to identify settings where they would benefit from transparent and trustworthy AI predictions, and to develop algorithms that satisfy the properties most relevant to their needs.

Requires deep interdisciplinary engagement with practitioners and stakeholders. Bring problem back to the theory, and establish the game-theoretic quantities that satisfy the properties most relevant to the problem. Then apply leverage score sampling to the special structure of the problem and yield efficient algorithms.

## 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. 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

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 regression 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 to make certain words more likely in generated text or add information in the Fourier domain of generated images. But modifications to the output can be detected and eventually forged, 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.

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 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.