Research Plans

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. My research focuses on the following goals:

1. **Explainable AI**: I design methods to more accurately interpret opaque models.
2. **Effective Algorithms**: I develop algorithms to solve problems with social impact.
3. **Responsible Use of AI**: I implement guard rails to prevent the misuse of AI.

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 the theory of boolean functions. Work in my area also requires 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 machine learning and theoretical computer science venues such as NeurIPS, ESA, and AAAI.

In the remainder of this document, I outline my research plans as they relate to the three overarching goals; describing my related work, future plans, and collaboration opportunities for each goal.

More Accurate Methods 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 interpret learned concepts, supporting the improvement and refinement of future AI algorithms.

The standard method of explaining AI predictions is to use game-theoretic quantities like Shapley values, quantifying how changes to the input features affect the model output. Shapley values are a particularly popular choice because they satisfy four desirable properties: null player, symmetry, additivity, and efficiency. However, Shapley values are defined over exponentially terms and are therefore computationally infeasible to compute exactly in general. In practice, Shapley values are estimated using heuristics approaches. Arguably the most popular and effective estimator for Shapley values is Kernel SHAP. Kernel SHAP exploits an elegant connection to linear regression, sampling from an exponentially large linear system to estimate the Shapley values. While effective in practice, Kernel SHAP samples heuristically and lacks theoretical guarantees.

My Prior Work

I developed a theoretically motivated algorithm for estimating Shapley values, Leverage SHAP, that outperforms Kernel SHAP in practice and offers strong non-asymptotic guarantees [MW24]. The starting point of my work was an insight from randomized linear algebra: regression problems can be effectively subsampled using statistical leverage scores. The benefit of leverage scores is that they quantify the importance of each data point in the regression problem. Instead of sampling from the exponentially large linear system via a heuristic weighting as in Kernel SHAP, Leverage SHAP, as its name suggests, samples according to leverage scores. This sampling technique is theoretically motivated and offers strong non-asymptotic guarantees. Modifying the standard leverage score analysis to several standard optimizations used for estimating Shapley values, I showed that Leverage SHAP can provably recover accurate Shapley values with almost a linear number of samples. Further, not only does leverage score sampling offer theoretical guarantees, but the theoretically motivated Leverage SHAP also outperforms even a highly optimized version of Kernel SHAP in practice.

While Shapley values are popular, they are not the only game-theoretic quantity that can be used to explain AI predictions. Instead of the efficiency property, we may want another application-dependent property. For example, in the context of a loan applicant, the 2-efficiency property ensures that the attribution of a composite feature like net worth is the sum of the attributions of the sub-features like assets and liabilities. If we replace the efficiency property of Shapley values with the 2-efficiency property, we arrive at the related Banzhaf values. Banzhaf values are simpler than Shapley values and have been shown to be more accurately computed in practice. I wondered whether I could apply the leverage score sampling technique to Banzhaf values. While Shapley values are known to be the solution to a linear regression problem, this formulation was only known for Banzhaf values in a restricted setting. In order to apply the leverage score sampling technique to Banzhaf values, I designed a linear regression problem for which the Banzhaf values are the solution [LWK$^+$24]. The resulting algorithm, Kernel Banzhaf, is substantially than the existing Banzhaf value computation methods. Further, because of the structure of the Banzhaf linear regression problem, Kernel Banzhaf offers even stronger non-asymptotic guarantees than Leverage SHAP.

Future Directions

My prior work establishes more efficient and theoretically motivated methods for explaining AI predictions with Shapley and Banzhaf values. However, there are still many more game-theoretic quantities that can be used to explain AI predictions in different settings. I plan to continue to apply my regression toolkit to design efficient algorithms for computing these game-theoretic quantities, furthering the transparency of AI predictions.

I plan to work with stakeholders to identify settings where AI predictions need to be explained for the trust of users. Depending on the setting, there are different desirable properties that give reason to game-theoretic attribution techniques. I will use my regression toolkit to design efficient algorithms that satisfy the properties most relevant to the stakeholder.

One different setting is graph data, where graph neural networks are used to make predictions. By design, the graph neural network manipulates the features of nodes in a graph and then passes relevant information to neighbors. In this way, the graph neural network uses graph structure to make more accurate predictions. Like in the tabular data setting, it is important to explain why the graph neural network made a prediction. Unfortunately, standard game-theoretic attribution techniques do not take into account the graph structure and so lose the ability to explain what the graph neural network is doing.

An alternative value is the Hamiache-Navarro (HN) value, which is designed for graph data. In fact, when the graph is the complete graph, the HN value recovers the Shapley value. Mathematically, the HN value is the limiting value of a series games, which can be thought of as repeated matrix multiplication. This connection to matrix multiplication suggests that the HN value can be computed as a gradient descent problem. I plan to investigate the structure of the HN value and design efficient algorithms for computing it. The matrices are exponentially large and naturally lend themselves to randomized linear algebra techniques. Like for Shapley and Banzhaf values, I plan to use leverage score sampling to design efficient algorithms for computing the HN value.

In general, there is a rich game theory literature to describe attribution techniques but game theorists interested in describing properties rather than computing these quantities. As AI becomes more prevalent, the game theory attribution is an existing toolkit that can be applied in an axiomatic way to explaining predictions. However, the existing algorithms are infeasible for large datasets or heuristically designed without guarantees. I plan to apply my regression toolkit to design efficient algorithms for computing these game-theoretic quantities, ensuring that AI predictions are transparent and trustworthy.

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.

Collaboration Opportunities

Related to game theory and graph algorithms, and regression.

Trustworthy Treatment Effect Estimation

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.

Treatment effect estimation is important problem. Often randomized control trials allow us to estimate the effect of a treatment. However, not always possible: some people may need the treatment more than others, or simply, the treatment has already been assigned and we can only observe the outcome. This setting is called natural experiments and is common in social good applications. One application is evaluating the impact of a nonprofit program. RORCO is an early childhood literacy nonprofit that provides books to children at their pediatrician visits. Collaborate with them to design estimator, try many complicated methods but no guarantee and each have different results.

My Related Work

Starting point is practical testbed for treatment effect estimation. Curated dataset and built benchmark for evaluating treatment effect estimators. From this empirical investigation, found that doubly robust algorithms give the best performance. However, in the literature, quite uninterpretable and guarantees are asymptotic. Designed a simple algorithm that is theoretically motivated and analyze its variance in the asymptotic setting, exactly. Find the performance is comparable to the doubly robust algorithms but with a simpler estimator. RORCO already used this algorithm to inform future program development.

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.

Future Directions

I am interested in designing user-friendly algorithms with understandable theoretical guarantees. Would like statement like with this many samples, can guarantee this level of accuracy. Further, simplified method can be used by stakeholders to understand the impact of their program. Many estimators stated in asymptotic sense without strong guarantees. Encourage development of effective and simple estimators. In framework, I'm interested in new estimators that draw on regression connection.

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.

Collaboration Opportunities Statistical estimator design

Causal inference

Distortion-free Watermarking for Responsible Use of AI

Powerful tools enable humanlike text and realistic images

While companies work to add guard rails, malicious actors can still use technology in any number of ways. For example, a malicious actor could claim that AI-generated text is their own or use AI models to generate images of fake events, causing confusion or even harm. Current approach is to watermark images and text so that content can be identified. However, most watermarking techniques are distortion-based, meaning they modify the output to embed the watermark which, in turn, can be detected and forged. For example, use red/green list of words to modify distribution of words in text so that with enough examples, the modified distribution can be recovered. For images, use Fourier patterns in images or in the latent noise of the image to embed the watermark. Again, detectable and forgeable with enough images.

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.

My Related Work

Designed distortion-free method for watermarking diffusion methods. The idea is to use a finite set of seeds for creating initial noise. Even if adversary can reconstruct the noise, the noise looks random and it gives no information about any of the other seeds because of the use of a cryptographic hash function. The challenge is storing all of the different possible noises and searching over them.

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.

Future Directions

Goal is to generate distortion-free method without the storage that scales. The idea is to use the context to get a seed in a robust and secure way. The technical ingredient is locality sensitive hashing.

LLM setting: Text is generated in an auto-regressive way by repeatedly generating a distribution over the next word. Prior work modifies the distribution, either in a global way or in a way that depends on the context. Either way, this is detectable with enough samples and therefore removable and forgeable. Not to mention the quality of the text is degraded. I plan to use an algorithm like SimHash to turn an embedding of the prompt into seeds. Then use a randomly selected seed with a cryptographic hash function to generate a random variable that is correlated with the noise. To detect the watermark, take each context and generate the correlated random variable for each seed and check for alignment with the observed text. By guarantees of SimHash, the seed is the same for nearby embedded contexts with high probability so the watermark can be detected with high probability. Can describe this process theoretically, and exploits streaming algorithms for distortion-free watermarking.

Vision setting: Similarly, store the information in the output. In this case, use prompt to embed and then SimHash to get many seeds, each seed is used to create the randomness in a different portion of the image. At detection time, we can caption the image to get a vector that aligns with the original prompt and then apply SimHash to get seeds. If any of the noise generated by the seed aligns with the reconstructed latent noise, conclude watermark. Again we have probability that depends on closeness of vector embedding and can vary hyperparameters to get desired level of detection. Also secure because the seeds are passed through a cryptographic hash function.

Dual combination of distortion-free and searchable. Streaming and randomized algorithms for speed up!

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.

Collaboration Opportunities at ETH Zurich

Randomized algorithms

Hashing and cryptography

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]. I have even explored board games through an algorithmic and complexity lens [WL20, Wit21], which sparked my interest in the field. Looking ahead, I hope to continue exploring new theoretical areas at Davidson College.

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.

[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.