Algorithmic Recourse

Kshitij Kayastha

Drexel University
kk985@drexel.edu

Abstract

With the rapid deployment of algorithms in high-stakes domains such as finance, healthcare, and justice, recourse has emerged as a critical tool in explainable AI. Algorithmic Recourse provides individuals with recommendations to reverse unfavorable outcomes, addressing the growing need for transparency and accountability in automated decision-making systems. My research investigates the design of recourse methods that are not only effective but also fair, robust, and grounded in ethical considerations.

Introduction

With the rapid deployment of machine learning models in critical domains and the major impact of these decisions on people's livelihood, a surge of recent work in responsible machine learning aims to make these models fair (Berk et al. 2021; Barocas, Hardt, and Narayanan 2019), transparent (Lakkaraju, Bach, and Leskovec 2016; Rudin 2019), and explainable (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Smilkov et al. 2017). A recent line of work within the explainability literature, termed *algorithmic recourse* (Wachter, Mittelstadt, and Russell 2018; Ustun, Spangher, and Liu 2019), provides individuals who received an undesirable label (e.g., one whose loan request was denied) with minimum-cost improvement suggestions to achieve the desired label.

Problem Formulation

Consider a binary classification task where the input space is $\mathcal{X} \subseteq \mathbb{R}^d$, and the label space is $\mathcal{Y} = \{0, 1\}$, where label 0 denotes an unfavorable outcome (e.g., loan denial) and label 1 denotes a favorable outcome (e.g., loan approval). Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the hypothesis class. A probabilistic classifier $h \in \mathcal{H}$ assigns a probabilistic estimate to an agent $x \in \mathcal{X}$, representing the likelihood of receiving a favorable outcome, i.e., $h(x_0) := \mathbb{P}(y = 1 \mid x_0)$. Using a threshold $\tau \in [0, 1]$, the probability estimates can be converted to binary labels $\hat{y} = \mathbf{1}[h(x_0) \geq \tau]$, where $\mathbf{1}[\cdot]$ is the indicator function. When a classifier h assigns an unfavorable prediction to an agent $x_0 \in \mathcal{X}$, i.e., $h(x_0) < \tau$ implying $\hat{y} = 0$, recourse aims to find a minimal cost modification to x_0 that results in a favorable prediction. This means finding an instance $x' \in \mathcal{X}$ such that $h(x') \geq \tau$. Let the cost of modifying x_0 to x' be measured by a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\geq 0}$, then the recourse problem is formulated as the following optimization problem (Upadhyay, Joshi, and Lakkaraju 2021):

$$\mathrm{R}(x, h) = x' \in \arg\min_{x \in \mathcal{X}} \ell(h(x), 1) + \lambda \cdot c(x_0, x). \quad (1)$$

Here, $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{\geq 0}$ is a loss function (such as square error or binary cross-entropy loss), and the parameter $\lambda \geq 0$ serves as a regularizer, balancing the trade-off between achieving the target prediction and minimizing the cost of altering x_0 to x'.

Let the *price* of a recourse x' with respect to a model h be denoted as:

$$J(x_0, x', f, \lambda) = \ell(h(x'), 1) + \lambda c(x_0, x') \quad (2)$$

My Contributions

My work addresses key challenges in algorithmic recourse, with a particular focus on enhancing its robustness and fairness.

Learning Augmented Robust Algorithmic Recourse

A major limitation in existing recourse methods is the assumption that the underlying predictive models remain fixed. In reality, models are frequently retrained to accommodate shifts in data or evolving deployment contexts, which can render prior recourse recommendations ineffective (Dominguez-Olmedo, Karimi, and Schölkopf 2022). To address this, Upadhyay, Joshi, and Lakkaraju (2021) introduced a robust optimization framework and an algorithm (ROAR) that ensures recourse validity under adversarial model shifts. However, such robustness often comes at a significant increase in price, resulting in overly conservative and expensive recourse solutions (Pawelczyk et al. 2023).

To overcome this trade-off, my collaborators and I developed a more practical approach that balances robustness with price-effectiveness (Kayastha, Gkatzelis, and Jabbari 2024). Instead of relying solely on worst-case assumptions, we adopt a learning-augmented framework (Mitzenmacher and Vassilvitskii 2020), which has gained traction in recent algorithm design for mitigating the pessimism of adversarial analysis. This framework allows the recourse designer to

incorporate predictions about how the model might change. While these predictions may be imprecise, our objective is to design recourses that perform well when the predictions are accurate (ensuring consistency), while still maintaining strong guarantees under worst-case deviations (ensuring robustness).

Performative Algorithmic Recourse

While our learning-augmented approach aims to reduce the price of robust recourse by leveraging predictions about future model changes, its effectiveness depends on the availability of such predictions. Most existing robust recourse methods handle model shifts by assuming small, adversarial changes and optimizing for worst-case outcomes (Upadhyay, Joshi, and Lakkaraju 2021; Nguyen et al. 2022; Nguyen, Bui, and Nguyen 2023). Although these methods improve the reliability of recourse under uncertainty, they require an estimate of how much the model might change. Underestimating this change can result in invalid recommendations, while overestimating it can lead to overly conservative and costly solutions (Pawelczyk et al. 2023).

To address this, we consider a setting where the model itself evolves as a result of recourse implementation. Rather than assuming specific knowledge of the degree of model shift, we propose a new framework, *performative algorithmic recourse*, that explicitly accounts for changes in the model induced by the recourse process. Drawing inspiration from Perdomo et al. (2020), we develop a gradient-based algorithm that anticipates distributional shifts resulting from recourse implementation and incorporates these effects directly into the optimization procedure. This approach allows us to compute recourses that are both more realistic and more effective in dynamic environments where user behavior can influence future model updates.

Robust Equal Improvability

While both our learning-augmented and performative frameworks address challenges arising from model dynamics, either by predicting or modeling how systems evolve, they primarily focus on improving the reliability and cost-effectiveness of recourse at the individual level. However, recourse recommendations can vary significantly in cost and effectiveness across demographic groups, raising concerns about fairness.

To address this, prior work has proposed group-level fairness criteria for recourse. For example,Gupta et al. (2019) suggest that individuals across different subgroups should face similar recourse costs, while Guldogan et al. (2023) advocate for equal success rates post-recourse. Yet, these fairness guarantees are generally framed in static settings and do not consider the evolving nature of real-world systems.

In practice, as mentioned earlier, as individuals act on their recourse recommendations, the resulting shifts in population-level behavior can alter the data distribution (Fonseca et al. 2023; Rawal and Lakkaraju 2020), prompting updates to the underlying models (Upadhyay, Joshi, and Lakkaraju 2021). Moreover, these updates can invalidate previous fairness guarantees or disproportionately affect some groups' ability to benefit from recourse.

To bridge this gap, we introduced a new framework for fair and robust algorithmic recourse. We proposed a novel fairness criterion, *robust equal improvability*, which requires that the proportion of individuals able to successfully improve their outcomes via recourse remains approximately equal across subgroups even in the presence of worst-case model updates.

Recourse under Model Multiplicity

Much of my work thus far has operated under the assumption that a single predictive model governs decision-making. However, in many real-world settings, multiple models with similar overall performance metrics, such as accuracy, may be equally valid and deployable. Despite this similarity at the aggregate level, these models can produce divergent predictions for the same individual, a phenomenon known as model multiplicity.

Recent research has started to address the challenges that model multiplicity poses for algorithmic recourse (Leofante, Botoeva, and Rajani 2023; Jiang et al. 2024). These approaches typically aim to ensure recourse validity across a set of plausible models. However, some methods are limited to specific model families or rely on averaging across models, which can obscure worst-case outcomes and lead to fragile guarantees.

To more directly address concerns of robustness and individual reliability, my current work takes a game-theoretic approach. We model the interaction between the recourse-seeking individual and a decision-maker who may deploy any one of several plausible models as a Stackelberg game. In this setting, we adopt a minimax formulation, seeking recourse solutions that remain effective even against the most challenging model within the considered set.

This work extends my broader research agenda by incorporating model uncertainty into the design of robust and reliable recourse.

Conclusion

My research aims to advance the theory and practice of algorithmic recourse by addressing critical challenges related to various aspects of recourse. Through the development of learning-augmented and performative frameworks, I have explored how recourse can remain effective even as models evolve in response to changing data or human behavior. I have further extended these ideas to group-level fairness by proposing the notion of robust equal improvability, ensuring that access to effective recourse remains equitable across demographic subgroups under distributional shifts. Most recently, I have begun investigating recourse under model multiplicity, where I adopt a minimax perspective to design recourse that are reliable even when multiple predictive models are plausible. Collectively, these contributions represent a step toward more trustworthy, actionable, and equitable algorithmic decision-making systems. I look forward to continuing this line of research and engaging with the AIES community to refine and expand these ideas.

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1): 3–44.

Dominguez-Olmedo, R.; Karimi, A.; and Schölkopf, B. 2022. On the Adversarial Robustness of Causal Algorithmic Recourse. In *39th International Conference on Machine Learning*, 5324–5342.

Fonseca, J.; Bell, A.; Abrate, C.; Bonchi, F.; and Stoyanovich, J. 2023. Setting the Right Expectations: Algorithmic Recourse Over Time. In *3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 29:1–29:11.

Guldogan, O.; Zeng, Y.; Sohn, J.; Pedarsani, R.; and Lee, K. 2023. Equal Improvability: A New Fairness Notion Considering the Long-term Impact. In *11th International Conference on Learning Representations*.

Gupta, V.; Nokhiz, P.; Roy, C.; and Venkatasubramanian, S. 2019. Equalizing Recourse across Groups. *CoRR*, abs/1909.03166.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024. Recourse under Model Multiplicity via Argumentative Ensembling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, 954–963.

Kayastha, K.; Gkatzelis, V.; and Jabbari, S. 2024. Learning-Augmented Robust Algorithmic Recourse. *CoRR*, abs/2410.01580.

Lakkaraju, H.; Bach, S.; and Leskovec, J. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684.

Leofante, F.; Botoeva, E.; and Rajani, V. 2023. Counterfactual Explanations and Model Multiplicity: a Relational Verification View. In *20th International Conference on Principles of Knowledge Representation and Reasoning*, 763–768.

Lundberg, S.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774.

Mitzenmacher, M.; and Vassilvitskii, S. 2020. Algorithms with Predictions. In Roughgarden, T., ed., *Beyond the Worst-Case Analysis of Algorithms*, 646–662. Cambridge University Press.

Nguyen, D.; Bui, N.; and Nguyen, V. A. 2023. Distributionally Robust Recourse Action. In *11th International Conference on Learning Representations*.

Nguyen, T.; Bui, N.; Nguyen, D.; Sue, M.; and Nguyen, V. A. 2022. Robust Bayesian recourse. In *38th Conference on Uncertainty in Artificial Intelligence*, 1498–1508.

Pawelczyk, M.; Datta, T.; van den Heuvel, J.; Kasneci, G.; and Lakkaraju, H. 2023. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. In *11th International Conference on Learning Representations*.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative Prediction. In *37th International Conference on Machine Learning*, 7599–7609.

Rawal, K.; and Lakkaraju, H. 2020. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In *Advances in Neural Information Processing Systems 33*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5): 206–215.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

Upadhyay, S.; Joshi, S.; and Lakkaraju, H. 2021. Towards Robust and Reliable Algorithmic Recourse. In *Advances in Neural Information Processing Systems 34*, 16926–16937.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *3rd AMC Conference on Fairness, Accountability, and Transparency*, 10–19.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2): 841–887.