General

Dear reviewers,

We were very grateful to receive such thoughtful comments from all of you. Of course, we were also very happy to get overwhelmingly very positive feedback from you — it has been truly rewarding to see that you all see value in this type of work.

This leads us straight to the point on SoK, that two of you made. We were both happy and somewhat surprised to see that two of you consider this as more of a genuinely novel contribution, than a systematization of knowledge. We did not think our work was novel enough, because a) we build on a very simple observation (see Reviewer vV4a's comments) and b) the algorithmic strategies we suggest are very simple. We do hope that we provide a novel perspective on Algorithmic Recourse, but we have done so merely by restating the existing framework in a way that explicitly addresses the notion of external costs that we propose. All of that being said, if the consensus is that this paper is novel enough for the "research paper" track, then we are happy to go for that instead. That is provided this is possible, of course, and the consensus also reflects the view of the PC.

We also have a general question with respect to prioritization of requested changes. When you've had a chance to read our individual comments on all of your individual reviews, could you let us know if you strongly disagree with us on any particular point? Upon also seeing other reviewer's comments, do you have a sense of what we should prioritize?

Once again, many thanks for your time and thoughtful feedback. It is really nice to see so much engagement.

Reviewer ebVS

Review

This paper presents an study of negative dynamics that happen when using algorithmic recourse. The authors provide a framework to study this problem and conduct an empirical evaluation on synthetic and real datasets.

PROS:

Novel angle on algorithmic recourse Characterization of negative dynamics and the situations in which they happen Thorough empirical evaluation Proposal of fixes and evaluations Honest discussion of limitations CONS

Paper content does not match an SoK. I do not understand why the authors choose to present this paper as an SoK. There is no systematization of knowledge. Instead, the authors provide new results on the area. I enjoyed reading this paper. The authors make a good job at explaining the problem and motivating why it is relevant contextualizing their contributions

with respect to the state of the art. The proposed treatment is simple, yet effective to show clearly the negative effects that considering recourse as static in time can have.

I only have two remarks:

The authors could do a better job at interpreting what the metrics they propose mean. While mathematically they make sense, I missed more intuition about why they capture the problems the authors are trying to highlight. In the countermeasure section, I would have liked to see more connection to the motivating examples given at the beginning of the paper. How would these countermeasures help in the case of the University? As a minor comment, in several places in the paper citations are treated as subjects of sentences (e.g., "presented by [4]"). This is non-desirable. I recommend to revise to make the authors the subject (e.g., "presented by X [4])

Response

Dear reviewer,

Thank you very much, we are happy to see that you enjoyed reading our paper.

Regarding your two main remarks:

1. Metrics: We think this is a fair point with respect to MMD. Here, in particular, we tried to make sure to rigorously lay out the proposed method(s). In our understanding MMD is a go-to measure in these kinds of setting, so we relied on that and a detailed description of what it entails. Do you have a specific suggestion as to how we can address intuition here? With respect to the remaining measures we think we do provide intuition, but if there is anything in particular that you were missing here please shout.
2. Linking the countermeasures back to the examples is a good idea. If after addressing other points we still have space, we will add a sentence or two. Does that sound good? Please see the general comment regarding prioritization.

With respect to SoK, please also see the general comment.

Reviewer vV4a

Review

This paper builds on a simple observation: when people successfully implement the actions that counterfactual explanations have suggested to achieve a different classification, such actions may induce concept drift if changes to the features do not result in a corresponding change in the outcome that these feature values are expected to predict. Or to put it differently, if acting on counterfactual explanations actually changes the observed relationship between X and Y,

then a model's predictive accuracy will decline as more people take these actions. The authors build on this observation to point out that the cost of concept drift is likely to affect other people, not just the institution using the model to make predictions. This could be true for two different reasons. First, an institution that finds that the accuracy of its decisions declines as more people make use of counterfactual explanations may cease to provide counterfactual explanations altogether. Second, if the institution recognizes that it needs to retrain the model to deal with drift, **the retrained model will likely move the decision boundary even further away from the people who have not yet attempted to make any changes, thereby increasing the costs they would face in seeking to achieve recourse.** Much of the paper is spent trying to formalize these dynamics and to evaluate how serious the problem is when using different methods for generating counterfactual explanations. The authors run experiments on a range of real-world and synthetic datasets, find that the problem is often quite significant across all of the methods tested, with some faring better than others. Finally, the authors consider a number of mitigation techniques, most premised on the idea that the best way to avoid these dynamics is to ensure that counterfactual explanations suggest actions that are in line with the true data generating process.

As I mentioned in the title field, I think this is a **very nice paper and is well above the bar for acceptance**. It's well motivated, it engages quite effectively with the existing and relevant literature, it formulates the problem in a very sensible way, and it does a wonderful job presenting all of the formal and experimental work in a very clear and accessible manner.

All that said, it also strikes me that the paper is built upon an observation that is not at all surprising, even if it is not one that is given enough attention in the literature on counterfactual explanations: **taking the actions suggested by a counterfactual explanation may not change the underlying property that a model is trying to predict**. For example, it should be obvious that it's possible to change the values of the features that are considered by a credit scoring model without changing the underlying likelihood of default. The reason for this is again obvious: most machine learning models are only learning statistical patterns in the training data; they are not learning the causal relationship between features and an outcome of interest. What is necessary to achieve recourse from a model (i.e., achieve a different classification) may be very different than what is necessary to actually achieve the quality that the model is try to predict.

The likely dynamics that fall out from this observation are pretty obvious: acting on counterfactual explanations can cause concept drift (i.e., alternating the observed relationship between X and Y); it will decrease accuracy due to concept drift will impose costs on the decision maker; the decision maker will likely try to find some way to reduce these costs, either by ceasing to provide counterfactual explanations or by retraining the model; retraining the model **will push the decision boundary further away from the initial people who have not yet attempted to achieve recourse**. It seems a bit odd to bring to bear so much machinery on this problem when the underlying issue is rather obvious.

Perhaps this didn't strike the authors as particularly obvious because they did not engage with the related work on strategic classification, which has focused on the fact that

preventing gaming of machine learning models is really "causal modeling in disguise": https://proceedings.mlr.press/v119/miller20b.html. While much of this work assumes that decision subjects might be acting adversarially (i.e., making changes to feature values in bad faith, knowing that such changes will not affect the underlying property that they are meant to predict), the main observations of this work apply equally well to this setting. Indeed, as Barocas, Hardt, and Narayanan point out, "**gaming can be a problem even when decision subjects are not acting adversarially. Job seekers may expend considerable effort and money to obtain meaningless credentials that they are told matter in their industry, only to find that while this helps them land a job, it does not make them any better prepared to actually perform it." https://fairmlbook.org/legitimacy.html.** Again, with this other scholarship in mind, it is not at all surprising that the authors conclude that"to avoid substantial domain and model shifts, the generated counterfactuals should comply as much as possible with the true data generating process".

I still think this work is quite valuable, especially because it does such a nice job demonstrating these dynamics with real-world data sets using a range of counterfactual explanation techniques. **But I would caution the authors against staking out such a strong position on the novelty of the initial observation that motivates much of this work.**

Requested Changes:

As I mentioned at the end of the review, I suggest that the authors adopt a more modest position with respect to the novelty of their work. In doing that, I also encourage them to take a closer look at the work on strategic classification and to try to link some of the observations in that work to the problem they are seeking to describe. I believe the paper would be strengthened significantly in drawing such connections and using these observations to better motivate and ground their work.

Response

Dear reviewer,

Thank you very much for the detailed review — it is really great to see that you have engaged with the paper so much.

Turning straight to your main point of criticism, not relating this to existing literature on strategic classification was indeed an oversight on our end. It should be possible to tone down the narrative a bit and point to the literature you have brought to our attention. In light of this, an interesting avenue for future research would be to test approaches to counterfactual generation that do explicitly incorporate causal knowledge (e.g. Karimi et al. (2020)), which we briefly reference in the paper. Since we are pushing the page limit (and also in light of the quick

turnover now), we do feel that it makes sense to merely point these additional considerations out, rather then attempt to address them in this paper. Do you agree with this?

One other tiny thing: we observe that the decision boundary does in fact "moves away from the target class", therefore making recourse even less costly to "future" recourse seekers. Of course, there is a catch: we would expect that recourse providers who are aware of such dynamics refuse to provide any recourse at all, so ultimately the cost to future recourse seekers is that they would be denied recourse. We tried to illustrate this through example I.2, but perhaps we need to make this even clearer?

Reviewer cxf3

Review

This paper examines the problems of model and distribution shifts which arise when individuals act on recourses received for the model. It argues for a departure from studying the problem of recourse for individual users toward a collective and longitudinal setting. The new problem is formalized with a number of evaluation metrics derived from related literature or developed by the authors. The paper further explores several algorithmic mitigation approaches to prevent the shifts while providing users with useful recourses. The paper supports its arguments with an empirical simulation using multiple base models, recourse models, and datasets.

The paper is very well-written. It might serve as a good foundation for future studies of the problem, especially since the authors plan to release their simulation framework. It might not seem like a typical **SoK paper in that it does not survey a body of work over an existing problem, but rather derives and grounds a new problem using prior literature and an empirical study.** To my mind (I am only tangentially familiar with the algorithmic recourse literature) the paper provides high pedagogical value.

I enjoyed reading the paper and believe it could be accepted as is. One particular omission which the authors might consider adding are plots showing how the shift metrics change over the course of the simulation.

When it comes to the out-of-sample error measure, I am wondering: to what extent do you expect the out of sample error increase to be a problem in practice? That is, what would the sample that is set aside in the experiments correspond to? Would the data distribution among users who did not yet receive recourse remain intact? Or might one assume that they'd also learn what they need to do to receive a positive outcome from the classifier through a backchannel?

Requested Changes

I would have liked to to see plots showing how the shift metrics change over the course of the simulation.

Response

Dear reviewer,

thank you very much for your review, we are happy to see that you enjoyed reading our paper.

With respect to your main points:

1. We considered showing the sort of chart you have in mind, but ultimately resorted to reserving this for the supplementary appendix that we point to in the paper. The reason we decided to omit these types of charts from the main paper, was that they only provide marginal additional information and are somewhat harder to read than the bar charts. In particular, we observed that over the course of the simulation the induced shifts gradually increase roughly proportionately for the different generators, until they level off. From the paper:

"Finally, it is also worth noting at this point that the observed dynamics and patterns are consistent throughout the course of the experiment. That is to say that we start observing shifts already after just a few rounds and these tends to increase proportionately for the different generators over the course of the experiment."

Currently, due to anonymization the online appendix only lives as a HTML document in our repo. We were planning to make it easily accessible through Github pages, once the paper is released. We can then provide a hyperlink to the relevant charts. Does this sufficiently address the issue in your view? If not we could consider replacing one of the bar charts with a line chart showing the changes over time.

2. Regarding the out-of-sample error, there are two different things going on here. Firstly, to address the issue of uncertainty around our proposed shift measures, we use a form of cross validation: we simply rerun the experiment multiple times for the sample sample and leverage the fact that individuals are randomly selected for recourse. This should not be an issue in practice, since it doesn't actually involve any data splitting. Secondly, though, to assess model performance we reserve a holdout set (this is probably what you are referring to). It's a good point. Our immediate suggestion would be to add an additional layer of cross-validation, this time with respect to data splitting. So the standard procedure of training on (K-1) out of K folds for each k. As a matter of fact, we might consider getting rid of the first CV procedure and simply replacing it with this one.

Regarding your point on SoK, please so our general comment.