We would like to thank the reviewers for their comments and suggestions. We will incorporate their feedback in the revised version of the paper.

Reviewers #1 and #2. As discussed in lines 125-135, our algorithms will look for the counterfactual explanations $\mathcal{A}$ and decision policies $\pi(\mathbf{x})$ that maximize the decision maker's utility rather than the individuals' best interest[1]. As a consequence, it is true that, compared to the non-strategic setting, the policies that are optimal in the strategic setting may induce some of the subjects to pay an additional (immediate) cost to change features in order to receive a positive decision, as shown in Figure 6c in Appendix E. However, we would like to point out that any subject who would have received a positive decision under the decision policy that is optimal in the non-strategic setting will still receive a positive decision under the decision policy that is optimal in the strategic setting after they best respond. Moreover, subjects who do pay an additional cost to change features will always increase their outcomes $P(y \mid \mathbf{x})$, as shown in Figure 7 in Appendix F.3, and this is likely to increase their individual utility in the long term. In the revised version of the paper, we will expand our discussion regarding the potential of our algorithms to favor the decision-maker at the expense of the decision subjects, in light of the results shown in Figures 6c and 7.

Reviewer #2 and #3. If our submission is accepted, we will make use of the ninth content page of the camera-ready version to bring the definition of α, the algorithmic boxes for Algorithms 1 and 2, and the discussion of the cost function estimation to the main text.

Reviewer #1. Under our problem formulation, a decision d is beneficial to the individuals who are subject to (semi)-automated decision making if $d = 1$ (e.g., an individual receives a loan) and a prediction $\hat{y}$ made by a machine learning model is positive if $\hat{y} = 1$ (e.g., an individual repays a loan). In this context, note that, rather than explaining predictions by machine learning models as in previous work, we pursue the development of methods to find counterfactual explanations for the decisions, as argued in lines 38-42. We will clarify this in the revised version of the paper.

We will expand our comparison with the existing literature and further discuss the necessity to distinguish between decisions and predictions, as argued by several authors in a series of recent papers [23, 25-27, 47, 48].

Reviewer #2. As noted by the reviewer, some of our assumptions are quite strong, however, we still think they do not nullify our contributions, especially given the paucity of work in the area. That being said, we are hopeful to relax some of these assumptions in future work.

We will fix the statement of Proposition 4.

Reviewer #3. The "strategic setting" refers to a scenario in which individuals who are subject to (semi)-automated decision making use knowledge, gained by explainability, to change their own features to maximize their chances of receiving a beneficial decision. In our work, we formally characterize this setting mathematically for counterfactual explanations.

A counterfactual is a statement of how the world would have to be different for a desirable outcome to occur [13]. In our problem formulation, the world are the feature values $\mathbf{x}$, the desirable outcome is the positive decision $d = 1$, and the statement is the counterfactual explanation $\mathcal{E}(\mathbf{x})$. Given an individual with initial feature values $\mathbf{x}$ who would receive a negative decision $d = 0$, the counterfactual explanation provides her with an example of a feature value $\mathcal{E}(\mathbf{x})$ under which she is guaranteed to receive a positive decision $d = 1$. We will clarify this in the revised version of the paper.

The problem formulation is new. Previous work on counterfactual explanations [13-15, 36-37, 39] has focused on explaining predictions, rather than decisions, and has not investigated the connection between strategic machine learning and explanations. The most closely related work is by Tabibian et al. [23] in the strategic machine learning literature, however, they have considered a setting where decision makers share their entire policies with the individuals subjects to their decisions rather than counterfactual explanations. In this context, please note that we have included further related work in Appendix A. If our submission is accepted, we will make use of the ninth content page of the camera-ready version to bring that content to the main.

[1] We did not explicitly use the wording social welfare or decision subject's utility, however, it is in the individuals' best interest to maximize their utility.