We thank the reviewers for their constructive and thoughtful comments. We begin with addressing concerns raised by several reviewers and then proceed with individual responses to each reviewer.

Relation to Perdomo et al.'s recent paper on performative prediction. Our paper differs from Perdomo et al. (referenced as [31]) in several important ways. [31] focuses on a setting where a predictive model at time t affects the input distribution at time $t+1$, a special case of which is users changing their features according to the model. For this setting, they give conditions for when a retraining procedure reaches an equilibrium. Their focus is exclusively on minimizing predictive loss (through empirical risk minimization), and the predictive model matters through its effect on convergence. They do not consider the quality of decision outcomes, and outcomes may be arbitrarily bad both in the equilibrium and throughout the retraining process.

Whereas Perdomo et al. care only about predictive accuracy, we care also about the quality of the actual outcomes associated with decisions, and study the tradeoff between decision improvement and predictive accuracy. This is a substantive difference that entails the need for assumptions that provide causal validity (see below). Another important difference is that we consider a one-time interaction between users and the model. Consider, for example, mortgage buyers, ICU patients, or first-time medical consultation (e.g., oncology, cardiology, psychiatry, screening tests). Our model is also relevant in settings with feedback coming in slowly, with models being intermittently re-trained, and where decision outcomes are consequential at each step of the retraining process. The main technical contribution is in extending the conventional risk minimization paradigm to account for decision outcomes.

Assuming covariate shift. Our work considers how decisions affect actual outcomes. This is a causal problem, and as such, requires assumptions that ensure causal validity (in our work this comes in through our use of propensity scores for reweighting evidence). The assumption of covariate shift (or its analogs, e.g., [34]) is made elsewhere in the machine learning literature [32,29,33,38,41,45,12], and although this is a strong assumption, it is reasonable for many applications. There exist important domains in which violations are sufficiently minor that this is a reasonable model, see Mueller et al. [29] and Peters et al. [32] for discussion (giving writing improvement as one example domain).

As some reviewers suggest, it would be interesting to quantify the effects of relaxing covariate shift on our framework. One possibility would be to assume Lipschitzness—in our case, that $p(y|x)$ changes smoothly with changes to $p'(x)$. Allowing for changes in conditional outcomes would affect the correctness of propensity weights, but can be accounted for by smoothly increasing uncertainty intervals, or equivalently, reducing the confidence τ in regard to improving decisions. Quantifying the relation between the Lipschitz coefficient and τ looks interesting for future work.

Reviewer 1 suggests that the covariate shift assumption contradicts the purpose of the framework, raising questions about feedback loops. We respectfully disagree. Undesired outcomes, as in the policing example, are the result of sample bias in data (missing observations) and of the inappropriate use of predictive tools for decision making. Our regularized predictor doesn't remove the problem with missing data, but nor is it adding to the problem; on the contrary, through this framework one can avoid decisions that are not supported by data. In policing, covariate shift is a strong requirement, but can be motivated through a suitable choice of x that includes relevant information (e.g., location, kind of policing, kind of community messaging), and while decisions are likely to affect the marginal $p(y)$, we do not see a clear reason for why this would change the conditionals $p(y|x)$.

Other individual responses:

[R1] Choosing η: If decisions x' are observed (and η is assumed to be independent of f), then fitting η to a sample set $\{(x_i, x_i')\}_{i=1}^m$ reduces to a unidimensional search problem. Similar assumptions on the decision model (often in the form of cost functions) are common in the literature on strategic classification as well as recourse.

Computational complexity: Let $c(\cdot)$ be the cost of a forward pass, then $c(R) = O(c(\nabla_f(x)) + c(\ell(x')))$. For linear f, g, h and $x \in \mathbb{R}^d$, we have $c(\nabla_f(x)) = d$ and $c(\ell(x'))) = kd$ where for quantile regression $k = 1$ and for bootstrapping k is the number of bootstrapped models. The cost of the alternating procedure follows accordingly.

[R2] Thank you for your encouraging and positive review.

[R3] Example in intro: Thank you for this comment, we will revise accordingly.

Relation to Perdomo et al. paper: Please see above. Note that we cite it as [31].

[R4] Empirical evaluation: As you note, the counterfactual nature of our setting makes validation challenging. Many works face this challenge, and lacking query access to arbitrary inputs, must resort (as we do) to some form of simulation. We have been careful to follow best current practices (e.g., [39]). Note that our approach does not require access to (nor the existence of) a true model f^*, and we simply use f^* to generate counterfactuals $y' \sim p(y|x')$ needed for evaluation.