We thank the reviewers for their kind comments, and for their consensus view that our approach of porting decision theory backed by behavior economics into classical ML is a promising research direction. We are also thankful for the reviewers' concrete suggestions on improving the draft, which we will incorporate in the final version of our work. Based on the requests of the reviewers, we have added an additional fairness baseline to compare with EHRM, using the reweighting approach in *Data preprocessing techniques for classification without discrimination, Kamiran et al.* When compared with EHRM$(.4, .7)$ from the main paper, we see that EHRM performs favorably across a variety of metrics.

AR1: *Human-aligned risk and fairness.* We agree with the reviewer that human risk measures are not necessarily fair. However, the innate loss aversion provided by CPT, ensures that HRM-learned models always avoid drastic losses, and consequently ensure that all subgroups do not suffer from huge losses. Whether this will lead to a "fairer" model is what we intended to explore with the experiments. In future work, we will theoretically study the connections between HRM and fairness, and combining CPT and Calibration (Kleinberg et. al. [2016]) to derive human-calibrated risk measures.

Connection to existing fairness literature. We thank the reviewer for pointing out the work of Agarwal et. al. [2018]. In contrast to EHRM, the authors' method requires access to an explicit set of protected attributes during training. Nevertheless, it is certainly an interesting question to see how the weights in EHRM are related to the cost-weighted framework of Agarwal et. al. We would also like to point out that fairness is one of several significant facets of our paper. Our primary goal is to introduce CPT inspired risk measures and study the consequences of its use within ML.

AR2: *Machine learning and human decision making.* As pointed out by the reviewer, there are two levels of decision making in ML: model selection when training a model, and instance prediction when using a model. These two kinds of decisions are very much related. In traditional ERM, a model is selected over others when per-instance predictions are more accurate on average. Our work explores the consequences of HRM in both settings: Sec. 2.1 and Sec. 4.3 discuss the model selection consequences of HRM; Sec. 5 explores its consequences on per-instance predictions.

Table 1: $a = .4, b = .7$ ensures EHRM weighting function to be close to the median estimate of the CPT weighting function given in Kahneman et al. [1992].

EHRM(.4, .7)	Kamiran et al.
.8766, .0057	**.8767 ±.0067**
-.0831 ±.0158	-.0875 ±.0212
.8453 ±.0293	.8396 ±.0390
-.0422 ±.0157	-.0518 ±.0253
-.0120 ±.0135	-.0165 ±.0177
.0861 ±.0034	**.0824 ±.0038**
.0422 ±.0157	.0518 ±.0253

Optimizing EHRM. We use the following iterative optimization procedure: $\theta^{t+1} = \theta^t - \eta \sum_{i=1}^{n} w_i \nabla_\theta \ell(\theta; z_i)$, where w_i are the weights obtained by reweighting the empirical risk using the weighted CDF given by CPT. Note that this approach is still a heuristic, and relies crucially on the assumption that minor perturbations in θ, don't change $w(F_n(\ell(\theta; z))$. Deriving provably optimal optimization algorithms for EHRM is an interesting open problem.

Novelty of Section 4. We would like to point out that (a) information weighted densities have been studied in information theory (Oliveira et. al. [2016]) and (b) effect of the probability weighting function on skewness of a distribution has been studied in finance in the context of portfolio allocations. However, we believe that our observations are novel, as (a) has not been studied in the context of CPT, (b) has not been studied on model selection, and these properties have certainly not been discussed in a unified way in the ML community. We will clarify this further in future versions.

Defn. 1. and L176. 1) The word "given" is inappropriate. $F(\ell)$ is the CDF induced by the data distribution. 2) HRM avoids drastic losses for all subgroups, including minority groups. The first paragraph of AR1 contains more details.

Experiments. We refer the reviewer to Appendix B (Fig. 5) in the supplementary material for separate plots of majority and minority performance of EHRM and ERM. We will add further figures of FNR with more fine-grained settings of a, b. The model configuration for the gender classification task is as follows: 3 convolutional layers (with number of output channels $(6, 16, 16)$ respectively, kernel size $(5, 5, 6)$ respectively and a 2×2 max-pooling on the outputs of the first layer), followed by two fully connected layers (the first has 120 hidden units and the second is the output layer with 2 output units); all activation functions are ReLU and all convolutional layers use stride 1. We thank the reviewer for bringing up the point of multiple hypothesis testing. We will correct our confidence intervals for this. As noted in Table 1 in the rebuttal, even a single fixed setting of EHRM compares favorably to existing fairness baselines.

AR3: *L126-128.* Diminishing sensitivity discusses how humans are less sensitive to certain changes of probability. Since $F(\ell) = 0$ and $F(\ell) = 1$ are "boundary" events (L124-L125), the inverse S-shaped probability weighting function implies that humans are more sensitive to changes of $F(\ell)$ when $F(\ell)$ is closer to 0 or 1 than when $F(\ell) = .3$.

Related Work. We thank the reviewer for pointing out *Fairness Risk Measures, Williamson & Menon* and for their suggestion to explore other risk measures such as CVaR. We will include a discussion of the paper in the related works section, and add experiments comparing EHRM and other risk measures.

Why CPT for surrogate loss minimization? It is the case that surrogate losses in ML are different from monetary gains and losses studied in behavioral economics. In particular, in ML, risk minimization only considers losses. However, what CPT captures is the characteristics of human's risk preferences when they make decisions under uncertainty. When applying CPT to surrogate loss minimization, we are assuming that an ideal ML model shares similar risk preferences as humans, e.g. avoiding drastic losses. We will make the connection clearer in the final version of the paper. As the reviewer has suggested, quantifying the alignment of HRM and human utility is a promising research direction.