We thank the reviewers for their helpful feedback. Below we address specific comments.

Reviewer #1: The reviewer's point about the clarity of the algorithm in Section 5 is well taken. Please note that some omitted details were relegated to Appendix E. We'd be happy to flesh out the algorithm's description in the body of the paper, as the reviewer suggests.

Reviewer #2: The reviewer's main concern has to do with the experimental methodology. We understand the comment that our methodology (generate instances in a way that there is an envy-free classifier with zero loss) is "slightly backward." We wish to clarify, though, that the reasoning behind this experimental design is that, if we generated instances as the reviewer suggests, we wouldn't be able to identify the envy-free classifier that minimizes loss, and, consequently, we wouldn't be able to measure how far our algorithm is from the optimum. Since there are no existing algorithmic benchmarks, we would argue that it's sensible to evaluate our algorithm by generating instances in a way that the optimal solution — i.e., the minimum loss achievable by an envy-free classifier — is known upfront, as we say in lines 281–283.

Let us now answer the reviewer's specific questions, using the given numbering; in our revision we will clarify all issues the reviewer listed.

(i) This is referring to expected utility. Please see lines 131–136. The word "overall" is confusing and will be deleted.

(ii) Exactly, envy is computed for each pair and then averaged over pairs. In Figure 3, "negative envy is replaced with 0, to avoid obfuscating positive envy." Please see lines 310–311.

(iii) Absolutely, these fractions correspond to α and β from Definition 1. The purpose here is not to make a technical connection to these parameters, though, but rather to make Figure 4 more concrete by giving examples of two points on the solid orange and magenta lines.

Reviewer #3: The reviewer notes that the "paper offers an interesting, original notion of envy-free fairness," that "the paper is well-written," and that "it offers nice technical results." On the negative side, the reviewer raises two issues, regarding the notion of fairness and the existence of utility functions. Since both points are rather terse, we weren't quite sure what specifically the reviewer is concerned about — we apologize if we misunderstood.

On the notion of envy-freeness: The reviewer writes that envy-freeness is "well-studied in other disciplines such as sociology, psychology and economics," so it seems that the importance of envy-freeness as a notion of fairnss isn't in question. Rather, if we understand correctly, the reviewer is questioning whether the insights from these other disciplines carry over to the machine learning domain in practice. While human-subject experiments are needed to definitively answer this question,[1] we note that there is a significant body of empirical work on envy-freeness in computational fair division [1], HCI [3], and behavioral economics [2]. The main insight is that people perceive situations where they are envious — i.e., those where they have a higher utility for someone else's outcome than for their own — as unfair; there is no reason why the same conclusion wouldn't hold in the classification setting, as it shares many of the same characteristics. We'd be happy to elaborate on this point in our revision.

On the existence of utilities: The reviewer writes that the paper "makes a very strong assumption of the existence of a particular form of utility function $u(x, h(x))$." We actually view this as a very mild assumption. All we're assuming is that each individual has a utility for each outcome, and the utility for a distribution over outcomes is the expected utility. Such utility functions, known as von Neumann-Morgenstern utilities, are the basis for much of the work in economics, decision theory, algorithmic game theory, and related disciplines.

References

[1] Y. Gal, M. Mash, A. D. Procaccia, and Y. Zick. Which is the fairest (rent division) of them all? *Journal of the ACM*, 64(6): article 39, 2017.

[2] D. K. Herreiner and C. D. Puppe. Envy freeness in experimental fair division problems. *Theory and decision*, 67(1):65–100, 2009.

[3] M. K. Lee and S. Baykal. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 1035–1048, 2017.

[1] There are preciously few such studies even with respect to the well established notions of fairness in machine learning.