We thank the reviewers for their time and feedback, which will help us to improve the clarity and framing of the paper. We are pleased that most of the reviewers commented positively on the importance of our chosen problem and the soundness of our results. We recognize that our submission has substantial room for improvement in 1) clarity presenting our results; 2) appropriately grounding our model in current applications; and 3) discussing and motivating the assumptions in our model. We are confident we can address these concerns in a camera-ready submission.

The goal of the paper is to study the brittleness of incentives in the context of the design of AI systems. As we note in the paper, this problem is well-studied as it relates to (incomplete) contracting in human systems and has received a lot of attention from the AI safety community. Our results apply to situations where an AI system is able to steer the environment towards undesirable states and the overall objective is complex, in the sense that it depends on many features of the world that are costly to measure or describe. Our results apply to negative externalities caused by misaligned proxy metrics in current AI systems[1]. We use the example of recommender systems in the broader impact statement. In this application, the relevant values are complex and recommendation behavior has been shown to greatly affect people's emotions and ideology. Our model justifies the iterative, flexible nature of designing these objectives, provides a mathematical account of some of the issues that these systems have run into, and lays the groundwork for improvements on the state-of-practice in metric design and maintenance.

Reframing and Pedagogical Example (R1, R4) We propose to reframe our paper to emphasize intuition and applicability of the work. We will present results in the context of a simpler, more intuitive model (linear U and convex C). This permits short proof sketches in the body of paper. we will move the general results to an appendix. We will complement this model with a running example motivated by algorithmic content recommendation. The attributes in this problem are as follows: 1) the ad revenue generated from user behavior; 2) the amount of time users spend on the website; 3) the quality of engagement (e.g., the proportion of clickbait recommended to users); 4) the diversity of a user's content; and 5) the overall community wellbeing. The constrained resource in this case is the user's attention. This will ground our model in realistic applications where attributes are continuous, utility (from the designer's point of view) is increasing in each, and it is clear that some of these features are prohibitively hard to specify.

Re: Human Model Assumptions (R1, R2) We will update the paper to provide more context and justification for our human model. Monotone, continuous utility functions are fairly standard in economic theory and modeling. Optimizing for these functions over sets of possibilities is the premise of consumer/producer theory. Utility functions that take into account multiple different aspects of the world (and hence analyze trade-offs) are known as multi-attribute utility functions. We can include a brief summary of relevant aspects of multi-attribute utility theory to justify our model. We will clarify that this is by no means the only way to model human preferences and is not a perfectly natural description of every situation (e.g. attributes with discrete values, bounded rationality). Despite this, it is easy-to-understand, fairly expressive, commonly used, and mathematically clean.

Re: Additive separability (R1) As a matter of clarification, additive separability, which is indeed a strong assumption, is only needed for proposition 1. We can reword the beginning of section 4 to make this more clear. In particular, none of the theorems in section 4 require additive separability—they need only the U and C satisfy the requirements of theorem 2, and in certain cases, are weakly convex/concave.

Re: Robot model and setting (R2) Our robot model describes an agent which provides updates to the environment yielding incremental improvement on a metric. In general, it makes no assumptions about the specific method—RL algorithms are examples of methods that incrementally improvement a metric in an environment (often in unpredictable ways). More than just providing examples of consequences of misalignment, we model how misalignment tends to lead the consequences in these sufficiently rich environments.

Re: Related Work (R2, R3) We thank the reviewers for presenting work relevant to our paper. Craig Boutilier et al.'s overall work on preference elicitation (including a specific paper [2] regarding elicitation of additively independent multi-attribute utility functions) and Christiano et al.'s paper on Deep RL from human preferences represent methods for an AI agent to effectively learn and update its objective function based on interaction with humans. We'll be sure to include these in the related works section of our paper.

Re: Implementability of Solutions (R3) R3 observes that computing sensitivity and minimizing impact are non-trivial problems. We agree, and do not mean to imply that these solutions are easy to implement. Instead, we hope that theoretical work like our submission can motivate and organize research on these important problems.

[1] https://arxiv.org/abs/2002.08512

[2] https://www.aaai.org/Papers/AAAI/2006/AAAI06-253.pdf