We thank the reviewers for their feedback. In particular, (R1, R3, R5) asked many specific and thoughtful questions, with (R1) marking questions by priority. Thank you.

We're glad that all reviewers agree that the paper is well-written and that side effect avoidance is an important AI safety problem. We are excited that (R1) thinks this work is ground-breaking, novel, and will spur further research, that (R3) is excited by our scaling of AUP, and that our results are considered strong (R1, R5) and significant (R1, R3).

**Experiments.** (R1): using multiple auxiliary reward functions performed the same or worse than $|\mathcal{R}| = 1$. (R1, R5) ask for more comparisons. We will include results for `DQN` and for additional auxiliary reward functions. At (R1)'s suggestion, we tried using the primary reward function as the auxiliary reward function. This condition achieved return and side effects comparable to `PPO`'s, as its attainable utility shifts did not correlate with side effects.

(R5) asks why we didn't compare to state reachability preservation [14] or to AUP with uniformly randomly drawn auxiliary reward functions over observations (like [25]). Unfortunately, neither approach is remotely viable in SafeLife. We estimate that there are billions of reachable states in any given SafeLife level. Accordingly, we're aware of a team trying to train reachability-preserving agents, but even the `append-still-easy` task was far too hard. If we generated reward functions by uniformly randomly drawing a reward for every state, the corresponding Q-functions would have extremely high sample complexity in an environment like SafeLife. We used a VAE because the encoder provided sufficient structure for quickly learning the auxiliary Q-function.

(R2, R4, R5) ask what distinguishes our work from [25]. The original AUP paper suggested that state reachability preservation avoids the same breadth of side effects as AUP in [25]'s toy environments. We show that unlike reachability preservation, AUP scales to SafeLife; as acknowledged by (R1, R3), we are the first to demonstrate compelling results on any complex side-effect avoidance environment. Furthermore, we showed that AUP competitively accrues reward while avoiding side effects, while equipped with a single auxiliary reward function which was learned *unsupervised*, whereas [25] drew several dozen auxiliary reward functions from the uniform distribution.

(R3, R4) wonder about scaling AUP to even more complex environments. We share their interest in this prospect. Realistic settings might have too many side effect opportunities for a supervised penalty to work well. We believe that the efficacy of AUP's unsupervised penalty term bodes well for even more challenging domains. (R3): we will add more discussion of the challenges AUP may face when scaling further, such as our assumption of a no-op action.

(R4) asks what the side effect score (defined lines 115-117) means. Roughly, if `AUP` halves `PPO`'s side effect score, then `AUP` disturbed half as many green cells. Disturbing a patch of green cells corresponds to about 4 additional side effect score. For a qualitative demonstration of the significance of a $46\%$ side effect score reduction, we refer (R4) to the attached `GIF` files for `append-spawn`.

**Training details.** (R1): agents were evaluated on their $N_{\text{env}}$ training environments. `AUP` had no data advantage – it trained for the standard 5M total steps. While `AUP`$_{\text{proj}}$ skips the 100K VAE exploration steps, it still learns its auxiliary Q-function for the first 1M steps. Given more training time, `AUP` does not overfit and increase side effects. To the contrary, when running three seeds from 5M steps (the paper's time limit) to 15M steps, `AUP`'s side effect score changed as follows: `append-still-easy`: $-24\%$, `append-still`: $-53\%$, `append-spawn`: $+17\%$, `prune-still-easy`: $-12\%$. We performed hyperparameter search for `append-still-easy`, and then applied the method successfully to the other three tasks without further tuning. "Random exploration" refers to a uniformly random exploration policy.

(R2): the hyperparameter sweep shows the average side effect score and episodic return for the *last* episode, while Figure 3's charts show a rolling average. The discrepancy was unintentional; we agree it is confusing and will fix it.

(R5): we used curriculum training (following the original SafeLife paper, [27]) because PPO fails to learn with just one environment. "Learning $R_{\text{AUP}}$" refers to the process of training the agent with respect to the AUP reward function defined in equation 1. Note that when training with respect to $R_{\text{AUP}}$, the auxiliary Q-functions are fixed, as they have already been trained. We will clarify the difference between auxiliary and AUP training in camera-ready.

**Novelty.** (R2) finds our work lacks novelty because we "take an existing technique and apply it to an existing environment". This misses our main contribution: demonstrating AUP's scalability, which is a crucial consideration in AI safety. As acknowledged by (R1, R3, R5), our work provides the *first* strong empirical evidence that AUP (and side effect measures more generally) can scale to any kind of complex environment.

**Clarity.** The prior work section and the SafeLife task explanations will be improved until they are as clear as the rest of the paper, and we will incorporate (R1, R5)'s suggested citations. (R2): we will make all code available. We are committed to providing the clearest possible camera-ready paper and look forward to refining the paper to (R1, R2, R3, R4, R5)'s satisfaction.