We thank all reviewers for their very helpful comments. We'll fix all typos and minor issues, and incorporate the suggested changes. Because of space constraints, we only focus on answering the reviewers' major questions below.

**Reviewer 1.** Our model is a standard Stackelberg security game (SSG), which forms the basis of many real-world applications in security domains. For example, in the Los Angeles International Airport, the SSG model is adopted to help decide allocation of canine units to terminals, and similar systems are adopted by the US Federal Air Marshal Service to deploy armed marshals to commercial flights [Jain-An-Tambe 2011 AI Magazine]. There is a large body of literature on SSGs with many model variants developed for specific scenarios [Tambe 2011]. We focus on the standard (and also the most general) model as the first work on attacker manipulation; the choice of model is in line with previous works on designing defender learning algorithms (e.g., [Blum-Haghtalab-Procaccia NIPS'14; Peng *et al.* AAAI'19]).

Our policy-based framework wraps the defender's learning algorithm as a sub-procedure, and allows *any* learning algorithm to be used as this sub-procedure as long as the algorithm learns by observing the attacker's best responses (e.g., algorithms in [Letchford-Conitzer-Munagala SAGT'09; Blum-Haghtalab-Procaccia NIPS'14; Haghtalab *et al.* IJCAI'16; Peng *et al.* AAAI'19]). The actual learning process is therefore abstracted as a reporting stage in our paper.

**Reviewer 2.** Binary search finds the optimal EoP $\xi$ within any desired precision $\epsilon > 0$ in time $O(\log(\frac{1}{\epsilon}))$. It terminates when we can locate $\xi$ in a small interval $[a, a+\epsilon]$. We believe this is a common approach for similar searching problems on a continuous range. It would be too demanding to seek the *exact* $\xi$ even in the theoretical sense, as it's unclear whether the optimal $\xi$ is a rational number to allow for a computationally feasible representation in the first place.

In the QR setting, the attacker is still aware of the "bounded rationality" of the defender. All assumptions are the same as the previous sections, except that here the defender is allowed to further randomize her commitment. In particular, when we say "perfectly/boundedly rational behavior", we refer to perfectly/boundedly rational behavior **in the truthful setting** where the attacker does **not** manipulate (we'll clarify this in the paper). When the attacker does manipulate, however, such "perfectly rational" behavior may turn out to be suboptimal and even worse than "boundedly rational" behavior. Intuitively, this is because the "perfectly rational" behavior falls into a fixed pattern for the attacker to exploit, whereas the "boundedly rational" behavior adds uncertainty in the fixed pattern and complicates attacker manipulation, in which case the attacker's old trick of making the fake game zero-sum would not work anymore.

We choose the QR model because it strikes a balance between the following two unaligned aspects of playing against attacker manipulation: 1) we want to discourage/punish attacker manipulation; 2) meanwhile we don't want the cost of 1) to be too high for the defender. The QR model on the one hand adds uncertainty in the defender's behavior that discourages attacker manipulation to some extent, while on the other hand it approximates the optimal strategy by choosing better actions with higher probabilities (i.e., the *softmax* function approximates the *max* function). We say that this is the rationality of QR (as in the **un**truthful setting) behind its bounded rationality (as in the truthful setting).

Stackelberg games are two-stage *extensive form games* (EFG). Technically, however, there's perhaps not much benefit of approaching our model as an EFG (as mentioned in Conitzer's paper, the defender's strategy space is infinite so there will be infinitely many nodes on the game tree), but it looks very interesting to study similar manipulation in EFGs.

**Reviewer 3.** We fully agree that our leader policy design can be viewed as a mechanism design problem and appreciate the reviewer pointing out the connection. Our work differs from other mechanism design problems in the underlying SSG that decides the players' utilities, where the mechanism designer plays as the defender and the attacker can lie about their payoffs. Our contribution lies in applying mechanism design ideas to this specific setting to tackle manipulation to defender learning algorithms. We'll add relevant discussions to make both the connection and the comparison.

For the QR model in Section 6, we refer the reviewer to our response to Reviewer 2. In the experiments, results of the "perfectly rational" behavior are represented by the **green** curves and labeld "SSE" (since SSE corresponds to the strategy of a perfectly rational defender *in the truthful setting*). Thus, the QR policy performs better than SSE in most cases, and only sometimes better than the optimal policy (**red** curves). The reason it can perform better than the optimal policy is that it allows additional randomization in the policy commitment (also see footnote 4 in the paper).

We introduce EoP because the *worst-case utility* is unable to distinguish the quality of many policies. As shown in Proposition 5, when playing against an attacker type whose payoffs are zero-sum to the defender's payoffs, no policy can achieve anything better than the *maximin utility* — essentially, it's a "mission impossible" to play well against such a zero-sum attacker type. As a result, if we take the worst-case utility as the criterion, the quality of all feasible policies would be hindered on this zero-sum attacker type, and hence be considered to be *no* better than the *SSE policy* (defined in Line 226); there will be no room for improvement against this criterion other than letting the attacker manipulate. This is unreasonable: *we cannot simply consider a policy to be bad just because it underperforms in some "mission impossible"*. The EoP is therefore proposed to adjust our measurement by taking into consideration also *the degree of difficulty* of the "missions" that is measured by the best achievable defender utility in the truthful setting (as the denominator of $\text{eop}_\theta$, Line 252). There is no additional assumption made about the attacker's abilities when EoP is defined. Each attacker type always aims at maximizing their absolute utility. We'll update the paper accordingly.