Reviewer 1. We appreciate R1's recognition of the novelty of our contribution to MARL and the potential impact on a range of problems. We address R1's two concerns below. **Regarding our chosen baselines**, we note that baselines we include represent three major existing categories: 1) policy gradient and actor-critic with discrete or continuous "give-reward" actions are direct applications of conventional RL (which have been applied to multi-agent incentivization in recent work (Lupu et al. 2020)); 2) LOLA is an archetype of second-order approaches; 3) Inequity Aversion (IA) draws domain knowledge from models in evolutionary biology and sociology to alter individual rewards. Hence we believe the existing baselines allow for fair benchmarking of our new approach. While Social Influence (Jaques et al. 2019) does not have open-source code, we can make an indirect comparison by noting that IA reaches a score around 250 by 1.6×10^8 steps (Figure 3a in IA (Hughes et al. 2018)), outperforming Social Influence that reaches score of 200 by 3×10^8 steps (Figure 1a in Jaques et al. 2019) in the original Cleanup map with 5 agents. Hence, the fact that LIO outperforms IA in our experiments implies that LIO compares favorably with Social Influence. We clarify that LIO technically does not involve a second-order gradient, as the gradients are w.r.t. separate parameters θ (policy) and η (incentive function). **Regarding scalability**, we clarify that the bi-level optimization does not necessarily imply difficulty in scaling up, because the learning of incentives is conducted in a *pairwise* manner: in equation (7) for a fixed reward-giver agent i, each term of the summation corresponds to the pair (i, j) for recipient $j \neq i$. Figure 1 shows that LIO attains the global optimum collective reward of 17 ($= 2 \times 10 - 3$) in the Escape Room game with $N = 5$ agents, out of which $M = 3$ agents are incentivized to cooperate despite penalties of -1 each. Scaling up to large populations poses new questions regarding population-level phenomenon (such as social norms) that modulate the impact of incentives; we leave this to future work.

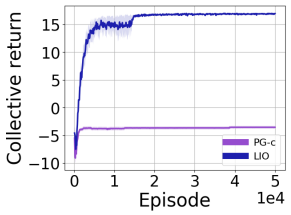

Figure 1: Escape Room ($N = 5, M = 3$)

Reviewer 2. We appreciate R2's positive feedback on our quantitative results and we are glad that our behavioral analysis of the learned incentive functions provided insight. We believe this work is a suitable contribution to the NeurIPS community, in addition to the broad area of multi-agent learning, as we tackle the open question of emergent cooperation from decentralized learning charted out in Hughes et al. (NeurIPS 2018) by building on general meta-gradient methods (Xu et al., NeurIPS 2018). In our revision, we will elaborate on the wide range of new research questions generated by our work, including theoretical analysis of dynamically-changing incentive functions, new population-level effects in a scaled up context, and adaptive ways to account for the currency of incentives. Regarding Figure 6b where the agent gives nonzero reward for "fire cleaning beam but miss" after 40k steps, one reason is that the agent's actual partner in training already converged to the behavior of consistently cleaning waste successfully (LIO in Figure 6a), so it may have "forgotten" the difference between successful and unsuccessful usage of the cleaning beam. As demonstrated more clearly in the Escape Room results (e.g. Figures 5b and 5d), this can be avoided by choosing a sufficiently large lower bound on the exploration rate by all agents, so that all agents pose the risk of deviating from cooperative behavior, which forces LIO to maintain correct incentivization.

Reviewer 3. We thank R3 for recognizing our contribution to the general class of opponent-shaping algorithms. We address each concern as follows. 1) Our definition of "decentralized" focuses solely on the inability to optimize social welfare directly, which is the crux of social dilemmas, and which holds regardless of access to the global state (e.g., Prisoner's Dilemma is fully observable). Hence our definition does not mention the degree of observability. 2) We explain in Appendix A that the coefficient α in the cost for incentivization is indeed an important hyperparameter and provide intuition for how to choose it in practice. We agree that a sweep over α can provide more insight. More broadly, we believe there is room to develop an adaptive scheme to trade off between small α, which allows time to learn the effect of incentives, and large α, which penalizes redundant incentivization. 3) In Figure 5d, by episode 50k, the cooperator still receives nonzero incentives between 0.5 and 1, but the winner's received incentives has noticeably converged to zero. 4) We became aware of Hostallero et al. (AAMAS 2020) after submission, and we believe there is a crucial methodological difference from our work. They use the temporal difference error of agent k's Q-function to modify the rewards of k's peers, so that those peer agents have incentive to take actions that lead to more favorable reward than average for agent k. This is a *passive* approach from agent k's viewpoint, and is closer to the method in Hughes et al. (2018), since each agent's original reward is modified by a hand-designed function based on other agents' performance. In contrast, a LIO agent *actively* differentiates through the recipient's learning step to update the learned incentive function, which it uses to change the recipient's total reward. We will include this in the revision.

Hostallero, D. *et al.* (2020). Inducing Cooperation through Reward Reshaping based on Peer Evaluations in Deep Multi-Agent Reinforcement Learning. *AAMAS*.

Hughes, E. *et al.* (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *NeurIPS*.

Jaques, N. *et al.* (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *ICML*.

Lupu, A. and Precup, D. (2020). Gifting in multi-agent reinforcement learning. *AAMAS*.

Xu, Z. *et al.* (2018). Meta-gradient reinforcement learning. *NeurIPS*.