Response about the significance and originality. Modelling the dynamics of multi-agent learning has long been an important research topic, but an n-agent setting where n tends to infinity has not been considered. All of the previous works focus on 2-agent settings and mostly use evolutionary game theoretic approaches (see the recent survey [Bloembergen et al., JAIR'16]). Our mean field theoretic approach is of fundamental difference from the evolutionary game theoretic approaches. As we explain in Introduction, the use of evolutionary game theory is inappropriate for n-agent settings, because, in principle, the number of equations required to model the entire population dynamics is proportional to the number of agents in the population. As n tends to infinity, analyzing or solving this system of equations becomes practically infeasible. In this paper, we show that by using mean field theory, only three equations are required to describe the dynamics of the whole population. A system of such small number of equations, as presented in Eq. 17, greatly reduces the problem complexity and makes the modelling tractable. Therefore, this paper introduces a new methodology to the modelling of learning dynamics in an infinitely large agent population, which is an emerging research topic given the growing interest in large-scale multi-agent systems.

The theoretical contributions of our work and the works [Mguni et al., AAAI'18; Mguni et al., AAMAS'19] mentioned by Reviewer 1 are very different. The works of Mguni et al. propose novel learning or incentive design methods, and prove that these methods will finally result in the convergence to efficient Nash equilibria in an infinitely large agent population. The actual process of convergence, however, is not formally described. This paper, to the best of our knowledge, is the first time to formally show the reinforcement learning dynamics, say, how the policies of individual agents gradually evolve over time, in an infinitely large agent population. In particular, the heart of this paper – a Fokker-Planck equation describing the evolution of the probability distribution of Q-values in an agent population – has not been reported elsewhere.

In this paper, we focus our attention on the population dynamics of an infinitely large agent population that use Q-learning. This is because Q-learning is one of the most important algorithms in reinforcement learning research and is the basis of a number of multi-agent reinforcement learning algorithms. Considering other learning algorithms will be an interesting and also plausible extension to our work.

We apologize that the above points should have been clearer. We shall highlight these points in the revised version.

Response about the experiments. The experimental study of this paper aims to illustrate and validate our mean field theoretic model. The games we select (prisoner's dilemma, stag hunt, hawk dove and choosing side) are typical matrix games that vary in the number, symmetry and efficiency of Nash equilibrium. This makes them good examples for illustration and validation, since they can be easily understood, but will lead to qualitatively different patterns of population dynamics. The nearly precise matching of the population dynamics derived from our model to those obtained from the agent-based simulations for each game type provides a clear and effective validation of our model. To further exhibit the strength of our model, we will find more complicated yet still understandable games to experiment on in the revised version. We will release our codes of the experiments if this paper get accepted.

Response to the questions raised by Reviewer 2. In each entry of Table 1, the first number is the payoff of the row player, while the second one is that of the column player. We shall include this specification in the revised version. The term $\gamma \max_{a' \in A} Q_t^{s',a'}(n_i)$ in Eq. 1 estimates the optimal discounted future payoff of player n_i under state s', after it plays action a under the current state s and consequently transits to the new state s'. For a matrix game, at a given time step t, agents play one round of the game. The row and the column players each takes one action simultaneously and receives an immediate payoff based on the joint actions. Then, the game ends. At the next time step $t+1$, the agents play a new round of the game. In other words, from time t to $t+1$, there is no state transition for an agent. Since there is no state transition at all, there is no need to maintain the term $\gamma \max_{a' \in A} Q_t^{s',a'}(n_i)$. Hence, it is a common practice to remove this term from the Q-value update function for matrix games [Gomes and Kowalczyk, ICML'09; Wunder et al., ICML'10; Kianercy et al., Physical Review E'12]. We apologize for the lack of explanation in the current version, and shall provide a detailed one in the revised version. We appreciate a lot for pointing out our typos and giving valuable suggestions on the language!

Response to the questions raised by Reviewer 3. The agent-based simulations are conducted on 100 agent populations each consists of $1,000$ agents. Agents play games strictly following the interaction scheme presented in Algorithm 1, and use Q-learning to update their policies. In Eq. 4 and 5, a should have been a_j. Eq. 5 holds for any valid value of η, which, by the definition of learning rate should be between 0 and 1. In Eq. 8, the series should be convergent, since the function $u(a, \mathbf{x}_t(n_j))$ is an analytic function. Given each element of the vector $\Delta \mathbf{x}_t(n_j)$ is between 0 and 1, we consider the second order and the higher order terms negligible. When $m, n \to \infty$, Eq. 8 holds. From Eq. 9, we can tell that the trajectory of each agent depends on its Q-values and is independent of who he/she is. Hence, we consider the trajectories of agents to be a function of Q-values in Eq. 10. We shall rewrite the left hand side of Eq. 10 to be $\mathbb{E}[\frac{dQ_i^{a_i}}{dt}](\mathbf{q}_t)$ for clarity. Given $n \to \infty$, the state of the population can be characterized by a distribution of Q-values in the population. Therefore, by deriving the Fokker-Planck equation that describes the time evolution of the Q-value distribution, we show in Eq.17 that only three equations are required to describe the entire population dynamics.