We would like to thank all the reviewers for their insightful comments. Their feedback has helped improve the paper significantly. Changes mentioned in our responses below have been incorporated in the revised version of the paper.

**Reviewer 1:** Regarding the contribution of the paper, our Level-1 theory of mind (section 2.2) was similar to Ref [23] except for the existence of the decay rate and having one set of parameters per subject for all conditions in the game, instead of one set per condition for each subject. However, our framework extends that work by serving as a basis for explaining the rationale behind human contribution by connecting it to conformity (section 2.1) as well as higher levels of theory of mind (section 2.3). The POMDP in Ref [23] can only explain the reward maximization aspect of human behaviour, while our general framework explains why human behaviour is optimal with respect to prosocial evolutionary behaviour and theory of mind. To the best of our knowledge, our paper provides the first formal definition of conformity, as well as higher levels of ToM for large groups. With regard to psychological interpretations, as higher levels of ToM are more complex, even with the same number of free parameters, better fit of a higher ToM does not guarantee its superiority over lower levels. That is not true for the opposite case. Better fit of lower ToM does mean superiority over higher levels. Therefore, while we might not be able to determine the exact level of ToM, we can suggest an upper bound for it. Also, normative models such as ours can be used in the field of computational psychiatry (for example see [1]). Specifically, the difference between ToM level, the prior, or the decay rate in patients and the control group is meaningful. Regarding the deterministic/nondeterministic policy of others and the agent itself, the POMDP model always generates a deterministic policy. In psychology/neuroscience experiments with reinforcement learning/Markov process models, stochasticity is added to the model's generated action by feeding the policy to a probabilistic function (e.g., see [22]). Similar to other classification models, this additional uncertainty does not change the prediction of each action (also mentioned in lines 190-192). It only changes the likelihood function of the model. Therefore, we don't need any new parameters to measure the accuracy of our model. However, if we want to make others' policy nondeterministic (according to the agent), we have to add at least one new free parameter. We will expand this part (especially the last paragraphs) in the final version of the paper.

**Reviewer 2:** Regarding the decay rate, higher decay rate (closer to 1) makes the previous observations and the prior more important. As the reviewer mentioned, this means that others' intention and consequently behaviour is more predictable and influenced by past events. Regarding the statistical tests, there is a good chance that rounds of each game, or even rounds of different games of the same subject are not independent from each other. As a result, to ensure the independence of samples in our statistical test, we used the average accuracy of each subject as one data point. We used the t-test because average accuracy is a continuous value and could be well approximated by a Gaussian distribution for all of our methods. We also ran the McNemar test, taking each round of each game of subjects as one data point. The results were in favor of our conclusions even more than the t-test (lower p-values) probably due to assuming a higher number of independent samples. Also, we compared our method to chance, with the permutation test. For both experiments the p-value was less than $0.001$. With regard to choice imbalance, choices are balanced in the consensus task (explained in detail in the original study [22]). In the VD, the number of free-rides was slightly higher ($56\%$). The balanced accuracy of our framework is $83\%$ significantly higher than the model-free method with $75\%$ balanced accuracy ($p < .001$). This means that our framework takes lesser advantage of the bias in the data than the model-free method.

**Reviewer 3**: 1-The reviewer is correct. We should have (and will in the final version) emphasized that this is a reasonable assumption only when others are not tractable due to the anonymity of actions (as in our experiments) or a large number of group members (as most of the real situations such as a jury). In that case, the subject assumes "an average group member" that generates actions because they cannot track individuals. 2- Each subject has their own set of parameters (including prior) in our framework. However, we assume that they "think" others have the same model as themselves. This simplifying assumption of 'you are essentially like me' is justifiable due to computational efficiency and anonymity of players. Moreover, this "false consensus" has been observed experimentally in humans (e.g. see [2]) 3- The concept of a decay rate is equivalent to giving a higher weight to more recent observations (samples). We used the decay rate instead of assigning a larger-than-one weight ($w \geq 1$) to the most recent observation for two reasons. First, to make our fitting methods computationally less expensive. Second and more importantly, we wanted to make our framework more aligned with the concept of decay rate or "leak" in psychology and neuroscience, which is used in decision making studies (e.g., see [3]). 4- We are sorry for the lack of clarity in this part. We will expand it in the final version of the paper. By assuming the same reward function for all players, the subject assumes that all other $N-1$ players choose the same action. Specifically, in round $t$, if the state (belief of the k-ToM agent) is $(\alpha_t, \beta_t)$, the agent assumes that all other agents choose $\pi^*_{k-1,t}(\alpha_t, \beta_t)$. 5- We totally agree with the reviewer that making the prior $Beta(1,1)$ is more consistent with the general definition of the prior in the framework. Our current choice, however, is more consistent with the interpretation of $\alpha_t$ and $\beta_t$ as previously experienced samples.

[1] P. Schwartenbeck and K. Friston. Computational phenotyping in psychiatry: a worked example. *eneuro*, 2016.

[2] M. Devaine and J. Daunizeau. Learning about and from others' prudence, impatience or laziness: The computational bases of attitude alignment. *PLoS computational biology*, 2017.

[3] M. Usher and J. McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 2001.