Evolving cooperation by useful delusions: subjective rationality and quasi-magical thinking

by Artem Kaznatcheev[1,2], Marcel Montrey[2], and Thomas R. Shultz[2,1]
[1]School of Computer Science and [2]Department of Psychology, McGill University

The role of subjective experience in human decision-making has been often ignored, or assumed to be too difficult to study. We combine ideas from biology, cognitive science, computational modeling, dynamic systems, and economics to introduce an evolutionary game theoretic framework for studying quasi-magical thinking and the distinction between objective and subjective rationality. The evolutionary fitness payoffs of interactions are given by an objective game which the agents are not aware of. Instead, each agent has a heritable subjective representation of what they think the game is on which they base their decision-making. The agents' minds consist of two probabilities: "will an agent cooperate with me if I cooperate/defect?" Since no agent can condition their behavior on the decision of another, any disagreement in these two probabilities is a misrepresentation of the world. We consider three different Bayesian updating techniques: purely rational, quasi-magical, and superrational. Given the genetic conception of the game and learned probabilities of cooperation, the agents act rationally on their beliefs (selecting the action that they believe to have the highest expected utility) .

We apply our framework to study the evolution of cooperation on k-regular random graphs. In this task, agents populate the nodes of a graph and engage in a pair-wise prisoners' dilemma interaction with their neighbors: an agent can maximize personal utility by defecting, but this decision minimizes social welfare and the whole population would be better off if everyone cooperates. We show that in highly specialized environments where cooperation incurs mild cost to the individual, the agents evolve misrepresentations of objective reality that lead them to cooperate and maintain higher social welfare. The agents act rationally on their delusions of the world but appear to behave irrationally when viewed by an external observer. In environments where the cost of cooperating is high, the standard prediction of universal defection is still achieved.

We consider two interpretations of the results. If the 'experimenter knows best' then evolution does not necessarily lead to accurate internal representations of the world, and the resulting misrepresentations are not always detrimental to individual or group performance. If 'evolution knows best' then pair-wise payoffs are not necessarily accurate depictions of the real game being played, and an overly reductionist account of the world can miss the social externalities inherent in decision-making. We discuss the connection between the evolved internal representations and inclusive fitness. Our simulations point out the importance of reasoning about internal representations and not just observed behavior when looking at economics or biology. Finally, we argue that our framework is a good foundation for combining evolution and learning since it captures the impossibility of defining an objective utility/fitness of memes if imitation dynamics are considered at the level of behavior.