Cooperate to Compete:
Coordinating Team Plans Across Time and Space

Michael Shum (mshum@mit.edu)

Abstract

Cooperation within a competitive social situation is natural part of human social life. This requires knowledge of teams and goals as well as an ability to infer the intentions of both teammates and opponents in order to coordinate or best respond respectively. We develop a model that jointly cooperates with teammates in order to compete against another team. In cooperating, agents can either assume teammates are as intelligent as themselves and plan with joint intentions, or assume that teammates are at lower levels like opponents are. We test predictions of this model in behavioral experiments using video-game like environments and then exploring augmentations to the model and environments for future work.

Introduction

From seeking promotions at a company to fighting over crayons in kindergarten, the social world involves competition amongst individuals. However social situations are not limited to competing against individuals and frequently involve cooperation with others to compete against other teams. These situations occur even early in childhood, in games like Tag or Keepaway. In these team games especially without explicit communication, cooperation doesn't emerge in a single moment but rather over the course of many actions. Individuals first have to identify friendly agents in uncertain situations and then demonstrate an intention to be teammates. Even after teams are known, humans still have to develop and execute detailed joint plans of action while inferring the plans of teammates[1].

Evaluating others is an essential skill in the social world. While we may make quick initial evaluations from physical features [2, 3, 4], we also make judgments about the friendliness of others based on their actions [5]. Hamlin et al found that 6- and 10- month old infants prefer individuals who help others to ones who hinder others, showing that even from an early age humans form abstractions about helpfulness and prefer individuals who help to individuals who hinder. This ability to infer someone's intentions from actions is called Theory of Mind [6], which humans use even at a young age to infer higher level concepts like goals [7, 8]. Baker et al., constructed models that probabilistically represent agents' desires and beliefs based on observing actions [9].

After observing other agents as friendly, it still isn't necessarily in our best interests to immediately seek to cooperate with them. One reason cooperation arises is multilevel selection[10], the idea that cooperation is due to cooperative groups being able to outcompete non-cooperative individuals[11]. If seeking to cooperate in a competitive situation, individuals have to form joint intentions and demonstrate the plan to teammates. This natural instinct to infer and evaluate social plans appears starting in early childhood[12, 13]. Children not only infer the goals of other agents but also execute complex plans to cooperate with others. In this work we aim to construct cooperative models that plan and coordinate like humans in competitive situations.

Naturalistic Games

We build games in naturalistic spatial environments that people play like video-games. This allows for intuitive emergence of actions following plans in order to reach strategic goals. These spatial environments are grids where players control individual agents. Simple grids like Figure 1 are a useful way to represent a variety of games since people intuitively orient themselves spatially and so form complex plans almost without any other knowledge.

In this work, we examine a single round of Tag between one team of two players and one team of one player. Each player controls the movement of one of the colored circles throughout the course of the game. On each turn players choose to move their circles into adjacent squares (not diagonal) or stay in the same spot. All players select an action during the same turn and all positions are updated simultaneously. If there are collisions between teammates they remain in the same place, while any collision between opposing players (moving into the same square, moving into each other's squares, or one player moving into a stationary player's square) counts as a tag and the end of the game.

After every turn the team that is "It" loses 1 point while the team being chased gains 1 point. Once the chasing team catches a single player the "It" team receives 10 points, the team being chased loses 10 points, and the game ends. As a result the game is zero sum with respect to teams. This structure is heavily based on games built by Kleiman-Weiner et al[14].

Model

Social Planning

We build a model of strategic planning that can form joint intentions assuming equally intelligent teammates, or varied lower levels of intelligence for other players. Agents know of the existence of teammates and share the rewards with them. At every step, agents select their action with a plan formulated under a presumption of each other player's intelligence. This model-based learn-

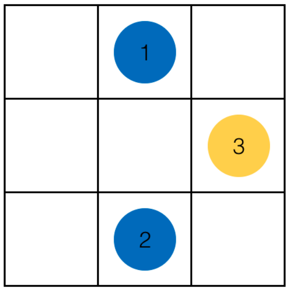

Figure 1: An example 3x3 state with three players.

ing generalizes well with multiple players as well as in new environments.

This work builds on classical formalisms of intention and joint planning from AI literature[15, 16] in addition to traditional reinforcement learning techniques[17]. Models in previous work do not handle uncertainty in a probabilistic way and so struggle with predictions about behavior.

Following the notation of De Cote and Littman[18], we construct stochastic games representing different children's games. A three-player stochastic game is represented as $\langle S, s_0, A_1, A_2, A_3, T, U_1, U_2, U_3, \gamma\rangle$ where S is the set of all possible states with $s_0 \in S$ as the starting state. If assuming equal intelligence, each agent chooses from a set of actions $A_a \times A_b$ constituting the joint actions of the team. Otherwise, an agent selects an action from its set of actions A_a. A state transition function $T(s, a_1, a_2, a_3) = P(s'|s, a_1, a_2, a_3)$ represents likelihoods of moving to new states given states and individual actions from agents. Reward for an individual player i is given as U_i. Additionally $0 \le \gamma_{game} \le 1$ is a discount rate of reward.

We define agents as attempting to maximize their joint utility, assuming other agents are doing the same. To represent this game-theoretic best response, we use the level-K formalism used in behavioral game theory with regards to the policies used by both teams[19, 20]. In a two-player game a level-K agent best responds to a level-(K-1) agent, which results in a level-0 agent. In our models, the level-0 agent is a randomly acting agent. This seems reasonable since the environment doesn't have specified goals, only to survive.

Joint Planning

If agents assume their teammates are at the same level they are in, the optimal action comes from treating the team as a single-agent[21]. As a result, they can construct level-K rollouts to identify the best team-action before marginalizing the actions of their teammate to identify their individual best action.

Accordingly, a randomly moving level-0 agent for a team with players i and j would have equal probability for any legal joint action for both players.

$$P(a_ia_j|s, k=0) = \pi_i^0(s) \propto \exp^{\beta Q_i^0(s,a_ia_j)}$$

$$Q_i^0(s, a_ia_j) = 0$$

With a level-0 agent defined, a level-k agent for players i and j on a team against player h on an opposing team can be recursively constructed in terms of lower levels.

$$P(a_ia_j|s, k) = \pi^G(s, a_ia_j) = \exp^{\beta Q_i^{(}k)(s,a_ia_j)}$$

$$Q_i^k(s, a_ia_j) = \sum_{s'} P(s'|s, a_ia_j)$$
$$(U(s', a_ia_j, s) + \gamma \max_{a_i'a_j'} Q_i^k(s', a_i'a_j')$$

$$P(s'|s, a_ia_j) = \sum_{a_h} P(s'|s, a_ia_j, a_h)P(a_h|s, k=k-1)$$

Here, player h is treated as part of the environment and so is described within $P(s'|s, a_ia_j)$. The maximization operator allows the joint agent to build the best-response to the level-(K-1) agent. Clearly this could be expanded to teams of any-sized n players against teams of similarly any-sized m players.

With a policy defined for the joint actions for a team by underlying $Q_i^k(s, a_ia_j)$, a single agent i on the team can marginalize out the actions of its teammate. $\pi_i^G(s, a_i) = \sum_{a_j} \pi^G(s, a_ia_j)$ and similarly for player j. These individual policies contain intertwined intentions that include an expectation for the teammate to reach certain states. This is a meshing of plans that is a key component of joint and shared intentionality[22, 23]. Crucially, agents here assume teammates are at the same level K they themselves are at.

Individual Planning

Agents can also assume teammates are at different levels than the ones they themselves are in. Notably, reward is still given if teammates achieve the goal. For this experiment, we assume all other agents are at level-(K-1).

A randomly moving level-0 agent then for player i only includes actions a_i.

$$P(a_i|s, k=0) = \pi_i^0(s) \propto \exp^{\beta Q_i^0(s,a_i)}$$

$$Q_i^0(s, a_i) = 0$$

Thus, an agent i doing individual planning at level-K with teammate j and opponent k constructs its policy with

$$P(a_i|s, k) = \pi(s, a_i) = \exp^{\beta Q_i^{(}k)(s,a_i)}$$

$$Q_i^k(s, a_i) = \sum_{s'} P(s'|s, a_i)(U(s', a_i, a_js) + \gamma \max_{a_i'} Q_i^k(s', a_i'))$$

$$P(s'|s,a_i) = \sum_{a_j,a_h} P(s'|s,a_ia_j,a_h)P(a_j|s,k=k-1) P(a_h|s,k=k-1)$$

The assumption that a teammate is one level lower is one that could be developed over time. In the future an optimal agent could infer the K-levels of other agents based on their actions over the course of the game and adjust to them. In our experiments, we utilized values of K-1 for both teammates and opponents and tested the self K to be either 1 or 2. This means agents expect all other players to be moving randomly, or best responds to an agent expecting everybody else to be moving randomly.

Behavioral Experiments

We constructed seven game states and asked 20 participants to pick where they would go in the next move as each player. Participants were given instructions that detailed the purpose of the game, goals of each player, scoring system, and dynamics of the environment. After seeing the state, participants were asked to select one of *Left*, *Right*, *Up*, *Down*, *Stay* as the next move for player 1, 2, and 3.

In comparing model predictions with human behavior, we tallied the count of each movement for each player for the state. We then created red heatmaps where each square's color intensity is proportional to the ratio of the movement count to all movements – the more people that chose a movement the redder the square that would be moved to. We also visualized the softmax policies for each model, with the probability of moving to a square determining the redness of that square.

Figure 2 shows the results of one model with individual planning and one model with joint planning. The individual model assumes it is level K=1, which assumes all other players are K=0. The joint model operates assuming the team is level K=1 and the opponent is level K=0. Globally we observe that both models capture the human data well, almost fully capturing the range of human decisions and generally capturing the distribution across actions as well. For all models we used a relatively high softmax β value of 7, as well as a γ discount rate of 0.9.

Between models, we note that the individual model places higher likelihood on the move it thinks is most optimal while the joint model places a small probability on moves that the human data shows. This can be seen more distinctly in the fourth row. Since our current model only does value iteration and state sizes grow exponentially as the board becomes larger, it was difficult to compute level K=2 heatmaps for all starting states. However, when conducting the experiment the most interesting human behavior was in starting state 2 (row 2); we will describe it in more detail as well as explore heatmaps for level K=2 models.

Human Ranges of K-Levels

We identified that K=1 models did not accurately capture the human sentiment to move down for player 2 in state 2 and built a level K=2 model to see if it was more accurate, shown in Figure 3.

For Player 1, both joint and individual models at level K=2 accurately captured the human intuition to move down or stay still. This intuitively makes sense since player 2 closes off the middle square.

Player 2 has the most diverse set of possibilities out of all squares shown. Human data shows participants equally preferred staying still, moving right, and moving down. Both level K=1 models strongly preferred moving right and slightly staying still, reflecting their understanding that player 3 moves randomly. However level K=2 models roughly equally weight staying still and moving down and place no weight on moving right. This captures the other human responses shown previously, which expect that player 1 will move down and prevent player 3 from escaping by moving left. It appears that different participants considered player 3 to be at different intelligence levels. Interviewed after their selections, participants who elected to move down said they didn't believe player 3 would be smart enough to move left. This would reflect a belief that player 3 is a level 1 agent.

Human data for Player 3 was almost entirely *Down*, with only two participants selecting *Left*. Similar to Player 2 this reflected the fact that most humans were thinking at a Level 2. However some participants selected *Left*, signifying that they were at Level 3. This would hope for Player 2 to move down, allowing Player 3 to escape through the middle.

It appears that humans operate as either level 2 and level 3 models. This decision may either reflect their ability to imagine steps ahead, a bias to underestimate the opponents, or a lack of time. Participants expressed an interest in playing out more moves in order to feel out other players' intentions. Some participants also said the more states they saw the less thought they tended to put into them due to the high initial cost of imagining scenarios. For future work it would be useful to explore the play of individuals over an entire game. This would allow us to more concretely identify what level individuals were playing at, as well as allow them to be more invested and accurate in their moves. The immediate feedback would be helpful in engaging participants.

We notice that humans never assumed opponents were moving randomly or that they were static. Most people projected what their possible range of actions were as opponents and did a best-response to that. However, humans interviewed said that they couldn't trust teammates to operate at the same level they were. This lack of trust in teammates was due to the zero-shot nature of the experiment, where humans weren't given any infor-

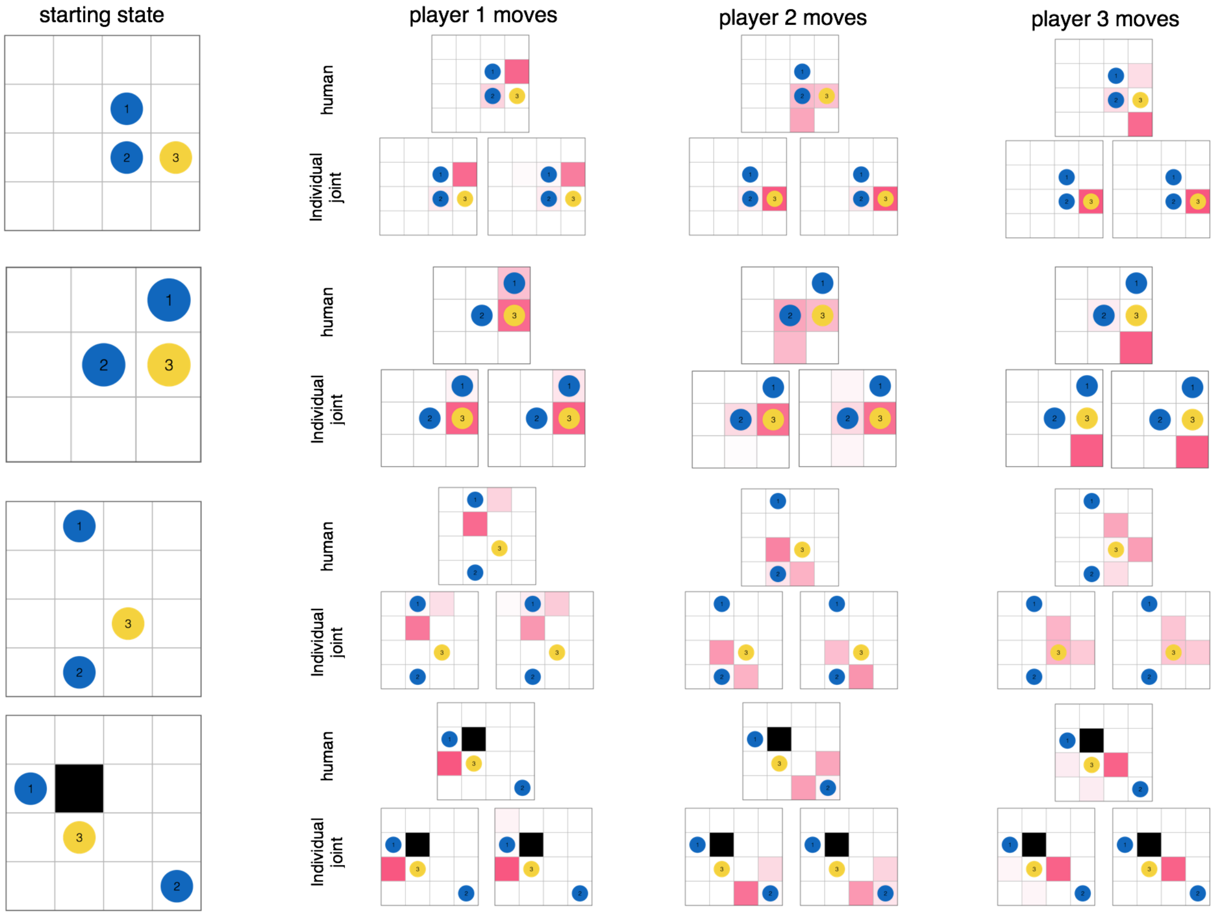

Figure 2: Participant tendencies and model tendencies at given states. Each row represents data starting from the state in column 1. The next three columns represent likelihoods of moving to particular squares for players 1, 2, and 3. Within each column, the top row represents human data. The left square on the bottom row is a model at K=1 assuming all others are at K=0. The right square is a model at K=1 assuming a teammate of K=1 and opponent of K=0.

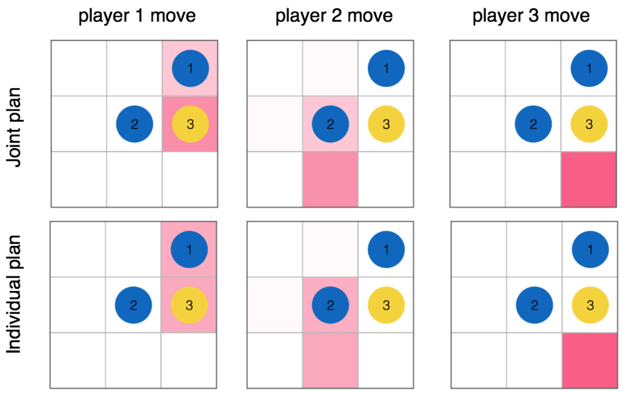

Figure 3: A level K=2 model's predictions for next moves in game 2.

mation about the other players. Upon observing moves, we hypothesize that humans are likely to gain trust in a teammate if they follow the joint policy at their own level-K. Notably a player might not gain trust in their teammate if the teammate's policy is a level-(K+1) since it might not be understood.

We note that a level-K policy only operates well in response to a level-(K-1) policy. Using Figure 3's state again, a level K=2 player two might move down with 60% probability which could allow a level-0 random player the opening to escape into the middle square. This is a key result from Wako Yoshida's Game Theory of Mind[24].

Future Work

Calibrating Models

Even when knowing teammates and goals, individuals have many parameters to infer in order to successfully accomplish these goals. The most important one identified is how intelligent one's partner is (previously mentioned as K level). Agents could be augmented to dynamically infer partner's K-levels and adapt their policies over the course of the game instead of fixing policies throughout the game. This same inference of K-level and adaptation could apply for opponents as well. A parallel for this might be having prior beliefs of opponents' intelligence or abilities from seeing them before and slightly modifying them based on observations during the game. A random level 0 agent also might not be the most accurate base model for some games, and can be changed depending on the environment and reward structure.

Optimizations to Planning

Since the state and action spaces grow exponentially with more players and larger boards, simple value iteration to calculate policies becomes intractable as a solution. For this experiment we were limited to board sizes of 4x4, and even those were frustratingly slow. We are considering using variants of TD learning such as Sarsa[17] to improve this speed. Other avenues might include Deep Q-Learning[25], which would be well-suited to our games due to the grid-world. Further, our current method calculates the full optimal policy for every state and every action, but we only need the best step from the state we are currently in. Many states are never seen and as a result we shouldn't need to calculate it. We also consider using Real-Time Dynamic Programming[26] for its efficiency in solving MDPs.

Other Games

In addition to this single round of Tag where the game ends upon catching an opponent, we will create other grid-world based games to explore human tendencies as well as our model's ability to model them. Grid worlds allow us to explore children's games well, so our plan is to model Tag (where It transfers upon being tagged), Freeze Tag (where It stays constant but players are frozen when tagged), Keepaway (where an object is being kept away from the other team) or a Lemonade Stand type game. Rewards are fundamentally different between these games – Tag and Keepaway involve a temporal reward that Freeze Tag and Lemonade Stand inherently don't, Tag has shifting team coalitions based on which players are It while Keepaway and Freeze Tag have established teams throughout. These simple variations lead to fundamental changes in individual behavior as well as team plans.

In addition to games with full knowledge, we could implement variations with limited visibility. Agents next to blocked squares may be unable to view where other players are due to an inability to view the rest of the board, and agents far away might not be aware of other players' exact locations or even existence. These uncertainties could be included in the models by averaging expected locations of agents around swaths of area, or have agents' locations be entirely unknown.

Inferring Teammates

In this experiment we played games where teammates were known, but it's possible to be in situations where teammates are not known or can change. We are interested in constructing models that can not only plan competitive joint policies with known teammates but can also infer what players are on the team.

A naive solution to this might be to construct policies for every possible team an agent could be on. Upon observation of actions, the agent can determine which constructed policy most resembles those actions. This could probabilistically update a posterior over which team the agent thinks it's on, which determines what policy it implements.

In addition to only observing actions, agents might update their priors on teams based on game situation. This reflects the fact that humans form teams not only based on loyalty but also based on environment. As an example, if we were playing Keepaway I might pass to you if you're far away and move to more open space so that we retain the object for longer.

Symbolically this might be represented as

$$P(T_m|a_i a_j a_k) \propto P(a_i a_j a_k|T_m)P(T_m)$$

where a_i, a_j, a_k are actions from all players i j k and $P(T_m)$ is the probability the agent is on team m. Here the prior $P(T_m)$ might all start as equal values across all teams, but could also be varied due to the world situation. $P(a_i a_j a_k|T_m)$ then could retrieved by indexing into the policies of the agent if it were on team m.

Establishing a Team

Teammates could be unknown, but there might be situations where agents don't know if they even need a team. We can consider this as the level above inferring teammates. This step would establish if one needs a team as well as who else should be on the team, and could be considered as creation of the priors $P(T_m)$ as well as construction of the variations of T_m. Further, agents could create long-term plans of forming and breaking alliances as well as potentially breaking up other alliances.

Here agents can also establish their norms when interacting with other players, since people don't solely seek to maximize reward. Individuals in real life may aim to maximize their own utility over others, even when participating in a team. In these children's games players may aim to be alive the longest in Freeze Tag or keep the object longest in Keepaway, sometimes for pride or reputation. This could cause them to join with players of roughly equal ability, or with many players of lower ability. Other norms for forming teams could emerge like ganging up against a heavy favorite to overthrow them or joining up with one in order to win, as well as showing mercy.

Demonstrating Teamwork

Having established that one needs a team and the players one would prefer on it, agents need some way to demonstrate this cooperative intent. There are fundamental differences to active cooperative intentions and lack of malicious ones that may be difficult for other players to realize, especially humans. Additionally if other agents are far away or unaware of one's own existence, the agent may have to move into their view to broadcast their intention. In order for this to closely mirror human intuition and best enable our model to cooperate with humans we may need to collect data on how humans infer intention from actions. This data could be used to train models in a model-free way or to inspire abstractions in model-based methods.

Discussion

In this work we constructed models that coordinate team plans with teammates in order to compete against other teams in a grid-based game of Tag. We compare the expected likelihoods for actions at given states for the model against human data and find that models at K=1 and K=2 fully cover the range of responses given. We note that joint planning is noisier than individual planning for K=1 as well as faster to calculate. While we were only able to construct results for low levels of K, with optimizations to our planning algorithm we expect to construct similar diagrams for higher K levels.

This work sets a strong framework around which we can explore abstractions to team-building dynamics in a variety of children's games. We hope to build the ability for AI agents in any multi-agent scenario where teams and goals are unknown to build coalitions and maximize reward in ways that humans intuitively do end-to-end. We are particularly interested in how models can learn norms at the team-formation level – humans can be altruistic[27], seek to collaborate with winners, but also can be drawn to the underdog[28]. How do these norms develop over time, and how can they be reconciled with concepts of loyalty? We aim to discover this in a model-based way through augmenting models with abstractions, as well as in model-free ways by learning from human data.

References

[1] A. Galinsky and M. Schweitzer, *Friend & foe: When to cooperate, when to compete, and how to succeed at both.* Crown Business, 2015.

[2] C. Herman, M. Zanna, and E. Higgins, "Physical appearance, stigma, and social behavior," in *Ontario Symposium on Personality and Social Cognition*, vol. 3, 1986.

[3] L. Winter and J. S. Uleman, "When are social judgments made? evidence for the spontaneousness of trait inferences." *Journal of personality and social psychology*, vol. 47, no. 2, p. 237, 1984.

[4] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall, "Inferences of competence from faces predict election outcomes," *Science*, vol. 308, no. 5728, pp. 1623–1626, 2005.

[5] J. K. Hamlin, K. Wynn, and P. Bloom, "Social evaluation by preverbal infants," *Nature*, vol. 450, no. 7169, pp. 557–559, 2007.

[6] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[7] C. L. Baker, R. Saxe, and J. B. Tenenbaum, “Action understanding as inverse planning,” *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[8] S. Liu, T. D. Ullman, J. B. Tenenbaum, and E. S. Spelke, “Ten-month-old infants infer the value of goals from the costs of actions,” *Science*, vol. 358, no. 6366, pp. 1038–1041, 2017.

[9] C. Baker, R. Saxe, and J. Tenenbaum, “Bayesian theory of mind: Modeling joint belief-desire attribution,” in *Proceedings of the Cognitive Science Society*, vol. 33, no. 33, 2011.

[10] D. G. Rand and M. A. Nowak, “Human cooperation,” *Trends in cognitive sciences*, vol. 17, no. 8, pp. 413–425, 2013.

[11] C. Darwin, *The descent of man and selection in relation to sex.* Murray, 1888, vol. 1.

[12] F. Warneken and M. Tomasello, “Altruistic helping in human infants and young chimpanzees,” *science*, vol. 311, no. 5765, pp. 1301–1303, 2006.

[13] K. Hamann, F. Warneken, J. R. Greenberg, and M. Tomasello, “Collaboration encourages equal sharing in children but not in chimpanzees,” *Nature*, vol. 476, no. 7360, pp. 328–331, 2011.

[14] M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, and J. B. Tenenbaum, “Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction,” in *COGSCI*, 2016.

[15] H. J. Levesque, P. R. Cohen, and J. H. Nunes, “On acting together,” in *AAAI*, vol. 90, 1990, pp. 94–99.

[16] B. J. Grosz and S. Kraus, “Collaborative plans for complex group action,” *Artificial Intelligence*, vol. 86, no. 2, pp. 269–357, 1996.

[17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* MIT Press., 1998.

[18] E. M. De Cote and M. L. Littman, “A polynomial-time nash equilibrium algorithm for repeated stochastic games,” *arXiv preprint arXiv:1206.3277*, 2012.

[19] C. F. Camerer, T.-H. Ho, and J.-K. Chong, “A cognitive hierarchy model of games,” *The Quarterly Journal of Economics*, vol. 119, no. 3, pp. 861–898, 2004.

[20] M. Costa-Gomes, V. P. Crawford, and B. Broseta, “Cognition and behavior in normal-form games: An experimental study,” *Econometrica*, vol. 69, no. 5, pp. 1193–1235, 2001.

[21] R. Sugden, “Thinking as a team: Towards an explanation of nonselfish behavior,” *Social philosophy and policy*, vol. 10, no. 1, pp. 69–89, 1993.

[22] M. E. Bratman, “Shared intention,” *Ethics*, vol. 104, no. 1, pp. 97–113, 1993.

[23] ——, *Shared agency: A planning theory of acting together.* Oxford University Press, 2013.

[24] W. Yoshida, R. J. Dolan, and K. J. Friston, “Game theory of mind,” *PLoS computational biology*, vol. 4, no. 12, p. e1000254, 2008.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[26] A. G. Barto, S. J. Bradtke, and S. P. Singh, “Learning to act using real-time dynamic programming,” *Artificial intelligence*, vol. 72, no. 1-2, pp. 81–138, 1995.

[27] H. Gintis, S. Bowles, R. Boyd, and E. Fehr, “Explaining altruistic behavior in humans,” *Evolution and Human Behavior*, vol. 24, no. 3, pp. 153–172, 2003.

[28] J. Kim, S. T. Allison, D. Eylon, G. R. Goethals, M. J. Markus, S. M. Hindle, and H. A. McGuire, “Rooting for (and then abandoning) the underdog,” *Journal of Applied Social Psychology*, vol. 38, no. 10, pp. 2550–2573, 2008.