The Impact of Other-Regarding Preferences in a Collection of Non-Zero-Sum Grid Games

Joseph L. Austerweil[†] and **Stephen Brawner**[*] and **Amy Greenwald**[*] and **Elizabeth Hilliard**[*] and **Mark Ho**[†*] and **Michael L. Littman**[*] and **James MacGlashan**[*] and **Carl Trimbach**[*]
Brown University
[*]Department of Computer Science
[†]Department of Cognitive, Linguistic, and Psychological Sciences
Providence, RI 02912

Abstract

We examined the behavior of reinforcement-learning algorithms in a set of two-player stochastic games played on a grid. These games were selected because they include both cooperative and competitive elements, highlighting the importance of adaptive collaboration between the players. We found that pairs of learners were surprisingly good at discovering stable mutually beneficial behavior when such behaviors existed. However, the performance of learners was significantly impacted by their other-regarding preferences. We found similar patterns of results in games involving human–human and human–agent pairs.

The field of reinforcement learning (Sutton and Barto 1998) is concerned with agents that improve their behavior in sequential environments through interaction. One of the best known and most versatile reinforcement-learning (RL) algorithms is Q-learning (Watkins and Dayan 1992), which is known to converge to optimal decisions in environments that can be characterized as Markov decision processes. Q-learning is best suited for single-agent environments; nevertheless, it has been applied in multi-agent environments (Sandholm and Crites 1995; Gomes and Kowalczyk 2009; Wunder, Littman, and Babes 2010), including non-zero-sum stochastic games, with varying degrees of success.

Nash-Q (Hu and Wellman 2003) is an attempt to adapt Q-learning to the general-sum setting, but its update rule is inefficient and it lacks meaningful convergence guarantees (Bowling 2000; Littman 2001). Correlated-Q (Greenwald and Hall 2003) is an improvement over Nash-Q in that, in exchange for access to a correlating device, its update rule is computationally efficient. However, there exist environments in which correlated-Q also does not converge (Zinkevich, Greenwald, and Littman 2005). Minimax-Q (Littman 1994a) converges to provably optimal decisions, but only in zero-sum Markov games. Likewise, Friend-Q and Foe-Q (Littman 2001) provably converge, but only to optimal decisions in purely cooperative and purely competitive games, respectively.

One significant shortcoming of the aforementioned multi-agent learning algorithms is that they define their updates in a way that makes assumptions about their opponents without actually factoring in their opponents' observed behavior. In a sense, they are too stubborn. In contrast, single-agent learning algorithms like Q-learning are too flexible—they simply adapt to their opponents without consideration of how their behavior will impact the opponent. What is lacking in these existing algorithms is the ability to *negotiate* a mutually beneficial outcome (Gal et al. 2004).

Algorithms have been designed that seek a best response against a fixed player and a mutually beneficial response against like players (Conitzer and Sandholm 2007; Bowling and Veloso 2002). Others attempt to "lead" a learning opponent to beneficial behavior (Littman and Stone 2001). In this work, we return to the investigation of the behavior of single-agent Q-learning in multi-agent environments.

Ultimately, a major goal for developing machine agents that act intelligently in multi-agent scenarios is to apply them to real-world problems. Humans are already agents that machine agents interact with in some current multi-agent environments (such as the stock market and online advertising auctions). Successfully expanding the scope of applications where multi-agent learning can be applied in the real world necessitates studying how these agents interact with human agents. A machine agent that interacts optimally against other machine agents, but not against human agents, is not likely to be effective in environments that include human agents. Further, one major goal of developing machine agents is for them to solve tasks in collaboration with human agents. Given the controversial nature of rationality assumptions for human agents (Kahneman, Slovic, and Tversky 1982), a machine agent that plans its collaboration by assuming the human agent will act rationally (optimally) is unlikely to be successful in collaborating with the human agent. Thus, in this paper, we investigate how human agents interact with each other, and how humans interact with fair and selfish reinforcement-learning agents.

Our work is inspired by results in behavioral game theory (Camerer 2003), where researchers have explored multi-agent decision-making in cases where each agent is maximizing a utility that combines their own objective utility and, to some lesser extent, other-regarding preferences that penalize inequity between agents. Our approach goes beyond earlier attempts to nudge agents toward more cooperative behavior (Babes, Munoz de Cote, and Littman 2008) and

instead provides a general framework that considers both objective and subjective rewards (Singh et al. 2010) in the form of other-regarding preferences. We investigate the behavior of this approach in machine-machine and machine-human interactions. Our main contribution is an exploration of how incorporating others' preferences into the agents' world views in multi-agent decision making improves individual performance during the learning phase, leads to desirable, robust policies that are both defensive and fair, and improves joint-results when interacting with humans, without sacrificing individual performance.

Experimental Testbed

Our experimental testbed included several two-agent grid games. These games are designed to vary the level of coordination required, while at the same time allowing agents to defend against uncooperative partners.

A grid game is a game played by two agents on a grid, in which each agent has a goal. See, for example, Figure 1, which is a 3×5 grid in which the two agents' initial positions are one another's goals: Orange begins in position (1,2), Blue's goal; and Blue begins in position (5,2), Orange's goal. We refer to grid positions using x-y coordinates, with $(1, 1)$ as the bottom left position.

One grid-game match proceeds in rounds, and each round consists of multiple turns. On each turn, the agents choose one of five actions (north, south, east, west, or wait), which are then executed simultaneously. In the most basic setup, agents transition deterministically, and there is no tie-breaking when two agents collide.[1] Instead, if their chosen actions would result in a collision with one another, neither agent moves. A round ends when either (or both) players move into their goal, or when a maximum number of turns has been taken.

As mentioned above, our grid games are specifically designed to prevent the agents from reaching their goals without coordinating their behavior. Consequently, one approach is for an agent to cooperate blindly with its opponent by simply moving out of the opponent's way, and hoping the opponent then waits for the agent to catch up. However, such strategies can be exploited by uncooperative ones that proceed directly to the goal as soon as their path is unobstructed.

To distinguish "unsafe" from "safe" cooperation, we devised a new categorization for strategies in our grid games. Specifically, we call strategies that allow for cooperation, while at the same time maintain a defensive position in the event that the other agent is uncooperative, **cooperative defensive** strategies. More formally, an agent's strategy is **cooperative** (C) if it is one that allows both it and its opponent to reach their goals, while an agent's strategy is **defensive** (D) if its opponent does not have a counter-strategy that allows it to reach its goal strictly first. A cooperative defensive (CD) strategy is both cooperative and defensive.

We now proceed to describe a sample set of grid games, and equilibria comprised of CD strategies (when they exist), to illustrate the kinds of interactions we studied. Our first

[1] It is a simple matter to vary these rules within our infrastructure, as future experimental design might dictate.

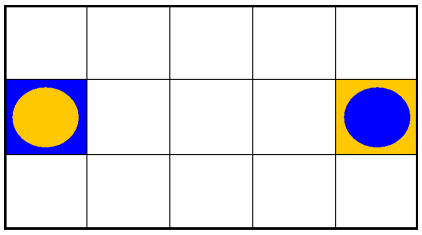

Figure 1: Hallway

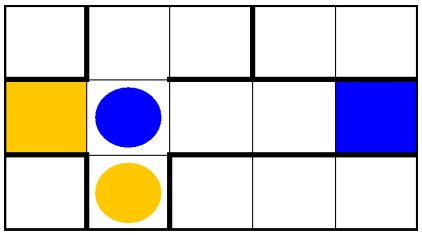

Figure 2: Intersection

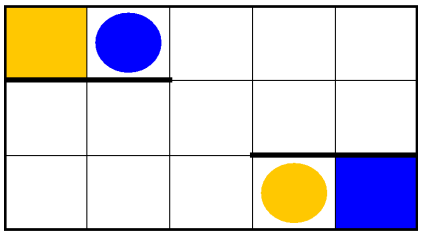

Figure 3: Door

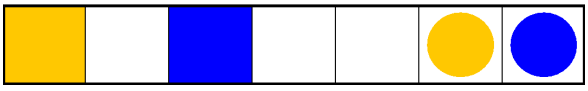

Figure 4: Long Hall

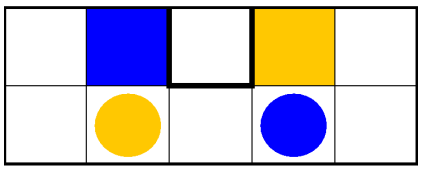

Figure 5: No Compromise

example, Hallway, is depicted in Figure 1. This game is one in which the agents can choose to coordinate, for example, if both agents agree upon a joint strategy where one agent moves along the top row and the other along the bottom, without interfering with one another. But, an agent could choose to "defect" from this joint strategy, by proceeding straight to its goal. There are CD strategies, however, that defend against this kind of non-cooperative behavior.

For example, if Orange moves south initially to $(1, 3)$ and Blue moves west to $(4, 2)$, Orange might choose to return and remain on its goal until Blue retreats to $(4, 3)$ or $(4, 1)$, at which point the players are equidistant from their goals, and both can reach them safely. This joint strategy is an equilibrium comprised of CD strategies, since Orange and Blue both remain in positions where they have the ability to block their opponents until they both have unobstructed equidistant paths to their respective goals.

The grid in Figure 2 (Intersection) requires Blue to defend against the possibility of Orange behaving uncooperatively, which it can achieve by squatting on the orange goal. Orange can then move to $(3, 1)$ where both agents are equidistant from their goals. Therefore, this game also has an equilibrium comprised of CD strategies for both players.

This equilibrium is not the shortest path, however. Purely cooperative agents in this game could adopt a joint strategy in which Blue moves east, while Orange waits a single step, before both agents proceed into their goals. This strategy profile is not defensive for Blue though, because it does not have the opportunity to observe if Orange will cooperate (wait) or defect (go north), and therefore cannot defend itself if Orange decides to head straight toward its goal.

Figure 3 (Door) is a grid that requires coordination to navigate through the narrow center space at $(3, 2)$. Any equilibrium comprised of CD strategies for this grid must be asymmetric, because it requires one agent to cede to the other agent the center cell. For example, if Orange chooses to cede that cell, it should step west into $(2, 3)$ while Blue steps south into $(3, 2)$. Then, Orange needs to step east back into $(3, 3)$ to prevent Blue from marching straight into its goal. Only when Blue agrees to step aside to $(2, 2)$ will they both be equidistant from their respective goals and in position to cooperate. This intricate pattern of first giving way to the opponent, and then forcing them to step around later represents an equilibrium comprised of CD strategies, since both agents are able to prohibit their opponent from reaching the goal first, but still leaves open the possibility for them both to reach their goals, cooperatively.

In the grid in Figure 4 (Long hall), Blue begins one step closer to its goal than Orange does. However, Orange can squat on the blue goal until Blue chooses to cooperate by taking one step back. If Orange can predict when Blue steps back, then Orange can take one step closer to its goal while Blue steps further away, in which case only Orange would reach its goal. The strategy that minimizes the risk to either agent requires that Blue wait one turn initially, while Orange moves toward its goal. These two strategies comprise a CD equilibrium.

Our last grid, shown in Figure 5 (No compromise), requires not only cooperation, but both agents must also exhibit trust for one another, or both agents cannot arrive at their goals at the same time. For example, Orange may sit on Blue's goal so that Blue can move to $(1, 2)$. Then, Blue must wait two turns before both agents are equidistant from the goals. If Blue defects and moves south into $(1, 1)$ while Orange moves south into $(2, 2)$, Orange still has the opportunity to go back up north to block Blue from reaching its goal. However, if Blue moves south into $(1, 1)$ when Orange steps east into $(3, 2)$, Blue will arrive at its goal sooner. Therefore, a trust spanning multiple rounds is required for the agents to effectively cooperate in this game.

No equilibrium in CD strategies exists for No Compromise. The game is like Door in that only one player can go move through the middle cell at a time. Unlike Door, however, it is not possible for the agents to simultaneously maintain a defensive position and to signal cooperation, because any cooperative move leads to an asymmetric situation in which the agents are no longer equidistant from their goals. As a result, after one agent cooperates, there is always an incentive for the other agent to defect, and there is nothing the cooperative agent can do to defend itself. Note however, that if Orange sits on the blue goal while Blue walks to $(1, 1)$ and then Blue cooperates by waiting for Orange to walk to $(1, 3)$, Blue's policy is CD. Still, Orange cannot respond in kind with a CD strategy; the aforementioned strategy is C.

Taking these five grid games as an initial testbed, we performed three studies: the first involved simulations of artificial agents playing against one another; the second pitted humans against other humans on Mechanical Turk; and the third, paired humans with artificial agents, also on Mechanical Turk. The remainder of this paper describes the results of these studies.

Machine-Machine Experiments

We carried out a set of simulation experiments with Q-learning in the grid games presented. For each grid game, we conducted 50 independent runs in which two Q-learners faced off. The agents' value functions were optimistically initialized with a value of 40 for all states and they used Boltzmann exploration with a temperature of 0.5. The discount factor was set to 0.9, and the learning rate to 0.01. To ensure that the state space was adequately explored, any state that is reachable from the initial state had some probability of being selected as the starting position for a round. Once either agent reached its goal (or 100 moves were taken), the round was terminated. There were no step costs beyond discounting, and rewards were 50 for reaching a goal.

We denote the outcome of a round using a pair of letters, where **G** means the agent reached the goal and **N** means the agent did not reach the goal. The first letter in the pair represents the agent's own outcome and the second represents its opponent's outcome. For example, **GN** is used to denote that the agent reaches its goal while its opponent does not.

As two Q-learning algorithms are not guaranteed to converge in self-play, we arbitrarily stopped the learning after 30,000 rounds, and checked the strategies learned. In spite of Q-learning not explicitly seeking outcomes with high so-

cial welfare, it very reliably identified cooperative strategies leading to **GG** outcomes.

Only the No Compromise game posed a challenge to the Q-learners. There, they tended to thrash about, finding a pair of strategies that work well together only to eventually discover that one of the agents has an opportunity to defect. The defection is destabilizing, so a new search begins for strategies that work well together. This result is not altogether unsurprising, because No Compromise is the only game in our testbed that does not possess a pair of CD strategies that constitute a Nash equilibrium.

In a second set of Q-learning experiments, we examined the impact of endowing agents with **other-regarding preferences**. That is, their rewards no longer depend solely on their own successes and failures—their objective rewards—but depend also on those of other agents in the environment as well. Standard reinforcement-learning agents, such as a Q-learning agent, seek to optimize their own objective reward signal—we call this preference the "selfish" preference because these agents are only concerned with their own outcomes.

When an agent has other-regarding preferences, however, it *believes* that this quantity is its true payoff. Previous work has shown, perhaps counterintuitively, that optimizing something other than the objective reward can sometimes lead agents to be more effective in their acquisition of the objective reward itself (Singh, Lewis, and Barto 2009). Indeed, we reach this same conclusion in our experiments.

Considering the four different outcomes in these games—**GG**, **GN**, **NG**, **NN**—there are 75 different possible preference orderings (allowing for ties). The selfish ordering that ignores the opponent's outcome is one of these orderings: **GG** $\sim$ **GN** $\succ$ **NG** $\sim$ **NN**. Nine of the 75 orderings are consistent with the selfish ordering, strictly preferring **Gx** to **Nx** for all **x**, and we are interested in which ordering leads agents to acquire the highest objective reward.

Of particular interest here is the **fair** preference, which we define as the objective reward of the agent minus 25% of the difference between its own and the opponent's objective rewards: $r_s = r_a - 0.25\,|r_a - r_o|$, where r_s is the agent's reward including its other-regarding preference, r_a is the agent's objective reward, and r_o is the opponent's objective reward.[2] By incorporating this fairness term, the agent strictly prefers the following ordering: **GG** $\succ$ **GN** $\succ$ **NN** $\succ$ **NG**. That is, the agent prefers making it to its goal as opposed to not, but it additionally prefers that the opponent only get to its goal if the agent itself does as well. To say it another way, a fair agent wants its opponent to win with it or lose with it.

Figure 6 shows the result of the selfish and fair agents playing against others of the same type in each of our test grid games. Of the nine orderings, only three (all variations of the fair preference ordering) achieve consistent cooperation in self play across all five grid games. Consequently, fair agents obtain higher total objective rewards than others. We also ran all nine preference orderings against one another. The average scores (across both players) in games involving a fair agent tend to be higher than the average scores in games not involving a fair agent.

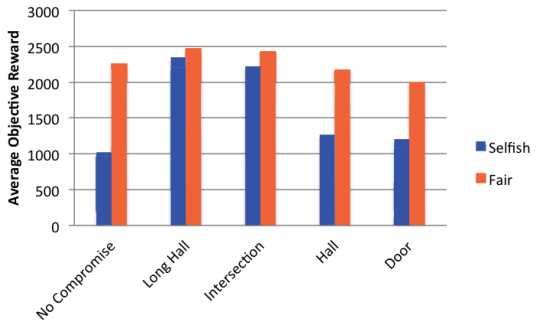

Figure 6: Average score in self play after 30,000 rounds.

We also analyzed the types of strategies learned by fair and selfish Q-learners after multiple simulations of various configurations. Interestingly, we found that Q-learners with fair preferences tend to find CD strategies more often, especially when paired with selfish agents.

In summary, Q-learners naturally learn to cooperate in the grid games studied, discovering equilibria comprised of CD strategies when they exist. Cooperation can be induced in other games by manipulating the Q-learners' other-regarding preferences to value the success of others.

In the remainder of this paper, we describe analogous experiments conducted with humans playing grid games. In those experiments as well, we were able to manipulate the rewards to favor fairness, and doing so induced more cooperation than otherwise.

Human–Human Experiments

We ran two studies in which human subjects played grid games. In the first, we recruited participants on Mechanical Turk to play the Hallway game against another Turker.

A total of 40 human participants were recruited via Amazon Mechanical Turk and were randomly paired (20 pairs) with one another to play as one of the participants in the Hallway game (Figure 1).[3]

Each participant began with an instruction phase that used a series of practice grids to teach them the rules of the game: arrow keys to move north, south, east, or west in the grid; spacebar to wait; both participants move simultaneously; when two participants try to enter the same square, their moves fail; and the round ends when either participant reaches a goal. Example grids demonstrated outcomes in which both participants reached a goal, and outcomes in which one did and the other did not. All transitions (including transitions that did not involve changing location) were animated so that participants could see that their actions registered. The instruction phase can be viewed at `http://goo.gl/SWme3n`.

[2] Other percentages would achieve the same result.

[3] One pair was not included in the analysis due to a technical error.

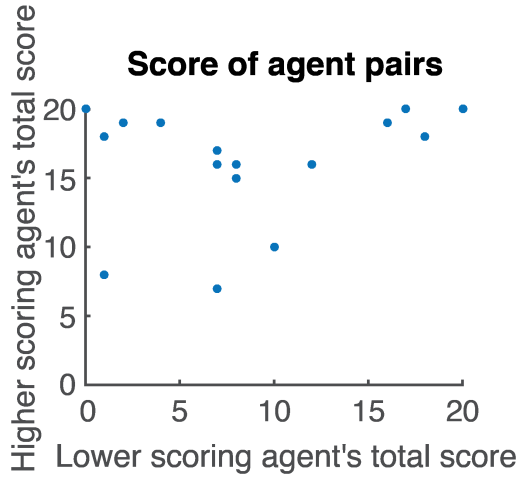

Figure 7: Human behavior in Hallway: The plot shows one point per pair, which represents the number of times at least one of the two participants scored.

After the instruction phase, the participants were paired up. Each pair played a match consisting of 20 rounds, which ended when either or both participants reached a goal, or when they had taken 30 actions without either reaching a goal. Participants received $2.00 as a base payment and a bonus of $0.10 for each round in which they reached their goal, regardless of whether the participant also reached their goal.

Participants were told they were playing against some other Turker.

An interface to view the actions taken by each pair of participants (and their feedback about the experiment) is available at `http://goo.gl/25IR5V`. Figure 7 summarizes this information in a single plot, which depicts a point for each pair that represents the number of rounds in which at least one of those two participants scored.

The plot shows a rich heterogeneity of behaviors. We broadly classified the outcomes into one of four patterns: trust (5/19), where participants reached their goals in nearly every round; alternation (5/19), where participants reached their goals on every second round (letting the other participant reach their goal in alternating rounds); surrender (3/19), where one participant reached their goal most rounds and the other participant tended to just get out of the way; and other (6/19), where some other pattern occurred (the majority of which were one participant reaching their goal most rounds and the other participant reaching their goal in about half of the remaining rounds).

Contrary to the Q-learning self-play results where all pairs converged to a cooperative strategy, only about one quarter of the human pairs behaved cooperatively. We propose two possible (and not mutually exclusive) hypotheses to explain this difference: (1) Our Q-learning agents were given 30,000 rounds to cooperate, whereas human pairs only interacted for 20 rounds, and (2) the learning strategies typically employed by humans do not tend to result in cooperation in this environment.

To better understand whether human agents could be induced to cooperate more reliably, and to understand what happens when a learning agent is paired with a person, we ran a follow up study in which people were paired with reinforcement-learning agents.

Machine–Human Experiments

Our first study on Mechanical Turk investigated pairwise interactions between human subjects, revealing a range of different outcomes. Our second study investigated the behavior of human subjects in Hallway when pitted against reinforcement-learning agents. The goal of this latter study was to investigate how human behavior might be influenced by the other-regarding preferences of a reinforcement-learning agent.

The experiment consisted of 19 participants who played against an agent, that is, a machine. The participants were told only that they were playing against another agent. The instruction phase was otherwise identical to the previous experiment.

There were two treatments in this study, defined by the two types reinforcement-learning agents, which differed in their subjective reward functions. Some were fair, while others were selfish. Given its reward function, an agent used value iteration to generate a policy against its current estimate of the human's policy. The humans' policies were estimated simply by counting the number of times an action was taken at each state. The estimated policy was then, at each state, an action with the maximal count.

In the selfish-opponent treatment ($n = 9$), the subject played against an agent with the objective reward function. In the fair-opponent treatment ($n = 10$), participants played against an agent with the fair other-regarding preferences. As in the human–human study, each participant was paid a $2.00 base pay and an additional $0.10 each time they reached the goal. They played a total of 20 rounds that each lasted up to 30 actions each. A viewer for the results is available at `http://goo.gl/nXm6IL`.

In line with our predictions, differences in other-regarding preferences led to differences in participant performance. In particular, fair-opponent subjects scored significantly more than selfish-opponent subjects ($t(9.0) = 2.37$, $p < 0.05$). Similarly, fair-opponent subjects tended to score consistently higher in objective reward than selfish-opponent subjects across the experiment (Figure 8). The agents themselves scored similarly between the two treatments.

Using a similar classification scheme as in the human–human experiments, we found that games played by selfish-opponent subjects resulted in trust (5/9), surrender (1/9), and other (3/9). In contrast, games played by fair-opponent subjects resulted consistently in trust (10/10).

Broadly speaking, the interactions between the humans and reinforcement-learning agents were either high in cooperation (many shared goals) and low in conflict (few collisions), or low in cooperation and high in conflict. Figure 9 plots the total number of shared goals against the number of collisions, averaged over all 20 rounds, and depicts two large clusters corresponding to these two types of interactions.

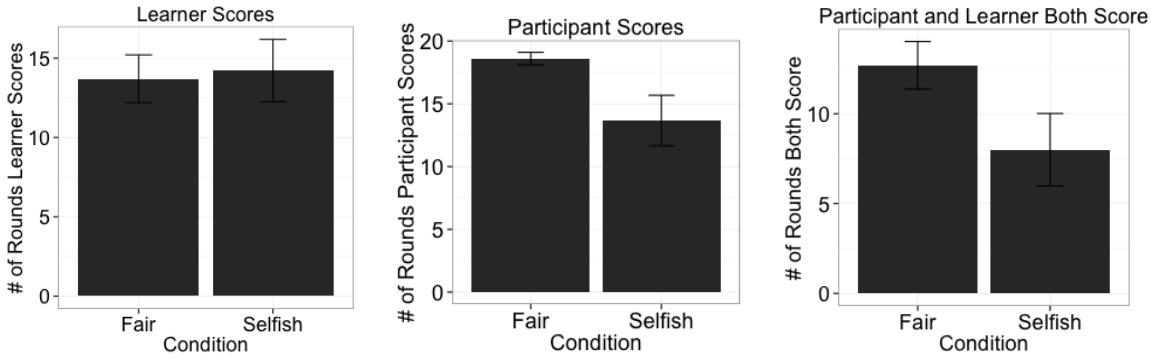

Figure 8: The impact of agent strategy on human–machine match ups in Hallway.

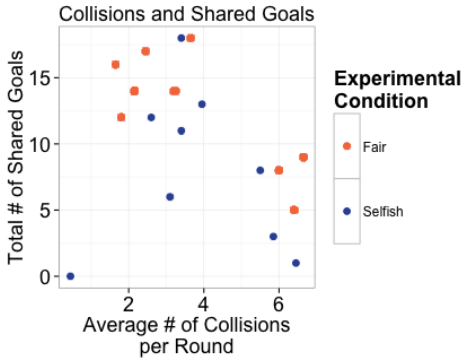

Figure 9: Average "collisions" (turns where both agents attempt to move to the same square) plotted against total number of shared goals in the Human–Machine experiment.

The first type of interaction suggests the emergence of a norm that does not require either agent to explicitly defend against the other, which implies that the agents had established some form of trust (top-left of Figure 9). The second one corresponds to a failure to agree on a joint policy that seamlessly allows both agents to reach their goals (bottom-right of Figure 9).

However, while the fair-opponent subjects overall tend to have more shared goals, the treatments split relatively evenly between the two clusters of interactions described.

Conclusions

In this work, we showed that introducing other-regarding preferences that favor fairness to reinforcement-learning algorithms can generate agents that have three positive qualities. They obtain higher rewards during the learning phase (Figure 6). When trained against a selfish learner, they are more likely to find CD strategies, guaranteeing a high level of objective reward against any future opponent and encouraging cooperation (Figure 9). And, they can improve the objective rewards of humans without decreasing their own objective reward (Figure 8).

Other multi-agent learning algorithms like Coco-Q (Sodomka et al. 2013) and Friend-Q (Littman 1994b) could also be used to promote cooperation, but will not behave as well when faced with an uncooperative opponent. In contrast, Q-learning, with other-regarding preferences that were different from but consistent with the objective rewards, is able to adapt its behavior to its observed opponent.

Q-learning, even when given useful other-regarding preferences, takes many thousands of interactions to converge to stable high-reward behavior. Our Amazon Mechanical Turk results showed that humans are able to converge to cooperative behavior much more quickly, resulting in high reward for all players. We showed that a model-based approach that explicitly estimates its opponent's policy could be endowed with other-regarding preferences and are able to quickly converge to high-reward behavior on timescales comparable to people. Further, these agents are able to converge to mutually beneficial behavior when interacting with people.

Ideally, we want communities of agents to discover, from experience, which other-regarding preferences are most appropriate to adopt within their population to achieve high objective reward. Future work will investigate how to address this challenge. One solution might be to incorporate "expert" algorithms (Crandall 2014; Megiddo and Farias 2005), which abstract the learning problem from learning about individual state-action pairs to learning which high-level strategy from a set of strategies to perform. In our problem, these strategies could be the set of strategies induced by different other-regarding preferences.

We reported on experiments involving (repeated) grid games played among humans, programs, and mixtures of humans and programs. As cooperation and "defection" are both aspects of our grid games, they naturally bear strong resemblances to the Prisoners' Dilemma. In numerous experimental studies of the repeated Prisoner's Dilemma, researchers find that people cooperate more often than game theory would predict (Camerer 2003). Possible reasons for this behavior include an overall preference for cooperation when one can determine that his/her opponent is also willing to reliably cooperate.

As in Prisoners' Dilemma experiments, our human experiments revealed a mix of behaviors, some of which were cooperative, and others which were not. Perhaps more interesting still is the fact that there was a clear distinction in behavior among the humans who played against the agents: humans whose agent opponents were fair were able to identify and exploit this condition readily. In future, more extensive, experiments, we intend to query the participants to try to determine whether or not they were interested in and/or able to identify one-another as cooperative. People who perceive machines as more predictable and having a preference for joint goal attainment might be more inclined to trust machines in joint tasks. This outcome would foster human–machine cooperation and could enable large increases in productivity by having machines share some of the workload in tasks that require more than one agent, one of whom is human.

References

Babes, M.; Munoz de Cote, E.; and Littman, M. L. 2008. Social reward shaping in the prisoner's dilemma. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems*, 1389–1392.

Bowling, M., and Veloso, M. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136(2):215–250.

Bowling, M. 2000. Convergence problems of general-sum multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 89–94.

Camerer, C. F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Conitzer, V., and Sandholm, T. 2007. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning* 67:23–43.

Crandall, J. W. 2014. Non-myopic learning in repeated stochastic games. *CoRR*.

Gal, K.; Pfeffer, A.; Marzo, F.; and Grosz, B. J. 2004. Learning social preferences in games. In *AAAI-04*, 226–231.

Gomes, E. R., and Kowalczyk, R. 2009. Dynamic analysis of multiagent Q-learning with e-greedy exploration. In *Proceedings of the 2009 International Conference on Machine Learning*.

Greenwald, A., and Hall, K. 2003. Correlated-Q learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, 242–249.

Hu, J., and Wellman, M. P. 2003. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research* 4:1039–1069.

Kahneman, D.; Slovic, P.; and Tversky, A., eds. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Littman, M. L., and Stone, P. 2001. Implicit negotiation in repeated games. In *Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, 393–404.

Littman, M. L. 1994a. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 157–163.

Littman, M. L. 1994b. Memoryless policies: Theoretical limitations and practical results. In Cliff, D.; Husbands, P.; Meyer, J.-A.; and Wilson, S. W., eds., *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, 238–245. Cambridge, MA: The MIT Press.

Littman, M. L. 2001. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 322–328. Morgan Kaufmann.

Megiddo, N., and Farias, D. 2005. Exploration-exploitation tradeoffs for experts algorithms in reactive environments. *Advances in Neural Information Processing Systems* 17:409–416.

Sandholm, T. W., and Crites, R. H. 1995. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems* 37:144–166.

Singh, S.; Lewis, R. L.; Barto, A. G.; and Sorg, J. 2010. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development* 2(2):70–82.

Singh, S.; Lewis, R. L.; and Barto, A. G. 2009. Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, 2601–2606.

Sodomka, E.; Hilliard, E.; Littman, M.; and Greenwald, A. 2013. Coco-Q: Learning in stochastic games with side payments. *JMLR Workshop and Conference Proceedings: Proceedings of The 30th International Conference on Machine Learning* 28(3):1471–1479.

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

Watkins, C. J. C. H., and Dayan, P. 1992. Q-learning. *Machine Learning* 8(3):279–292.

Wunder, M.; Littman, M.; and Babes, M. 2010. Classes of multiagent Q-learning dynamics with epsilon-greedy exploration. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning (ICML-10)*, 1167–1174.

Zinkevich, M.; Greenwald, A. R.; and Littman, M. L. 2005. Cyclic equilibria in Markov games. In *Advances in Neural Information Processing Systems 18*.