A Brief Overview of Social Dilemmas: From Social Sciences to Multiagent Based Simulations

Rafael M. Cheang[1,2], **Anarosa A. F. Brandão**[1], **Jaime S. Sichman**[1]

[1]Laboratório de Técnicas Inteligentes (LTI)
Escola Politécnica (EP)
Universidade de São Paulo (USP)

[2]Centro de Ciência de Dados – Universidade de São Paulo

{rafael_cheang, anarosa.brandao, jaime.sichman}@usp.br

Abstract. *Social dilemma is the name given to a set of games with conflicting individual and communal incentives and characterized by the existence of deficient equilibria. This area of research has for a long time drawn the attention of the social sciences, but not as much from the multiagent community. This paper aims to introduce this multidisciplinary area of research to computer scientists and engineers with a particular interest in multiagent systems. To this end, we present an overview of how the field has evolved over the last 40 years that includes: an introduction, key learning points coming from the social sciences, how social dilemmas are being simulated in multiagent Markov game settings, and the road ahead; difficulties and perspectives.*

1. Introduction

Great attention has recently been given to reinforcement learning (RL) algorithms and their prowess to learn from experience in complex environments. Milestones such as beating the best human players in competitive games[1] with huge state spaces, such as Go [Silver et al. 2016] and chess [Silver et al. 2018], are just a glimpse to show how much AI, and especially, RL, can help humans to solve problems that involve decision making under uncertainty.

Less attention, however, has been given by the multiagent systems community to an important subset of games that are not solely competitive but rather mixed motive. This collection of games, often called social dilemmas, has drawn considerable interest from social sciences‘ scholars [Dawes 1980] once it maps well into real-world social problems.

The importance of solving this class of problems is epitomized in one sentence by Dafoe et al.: ”There are multiple problems of global cooperation with stakes in the trillions of dollars or millions of lives, including those of disarmament, avoiding nuclear war, climate change, global commerce, and pandemic preparedness” [Dafoe et al. 2020].

However, dealing with these problems is not trivial. The mixed incentives nature of social dilemmas poses difficulties in guiding decision-making by pulling incentives to different and opposed directions: players are exposed to a short-term incentive to act greedily, which in turn, may compromise a long-term incentive to cooperate that would be beneficial to all involved.

[1]Also known as zero-sum games.

This paper aims to turn the attention of the multiagent systems community to this non-zero-sum set of games that we encounter with some frequency in our everyday lives. To achieve our goal, we first introduce the idea behind social dilemmas and how research on the domain progressed, especially through the 80s and 90s, based on a high-level abstraction of the problem — a matrix and a reward function. We then follow with a short review of studies based on a lower-level formalism that uses modern RL algorithms and pixel environments. We present the differences between models, difficulties encountered when switching from high-level to low-level formalisms, and finish by discussing some open problems in social dilemmas.

2. Delving deeper into social dilemmas

A social dilemma may be described as a game that exposes its players to simultaneous and conflicting cooperative and competitive incentives. The game is cooperative in the sense that a high collective reward can be achieved by mutual cooperation, and it is competitive because every player has an individual incentive not to cooperate, either by fear of the opponent not cooperating or greed. Every social dilemma is marked by at least one *deficient equilibrium* [Kollock 1998]; selfish incentives lead participants to a state where no individual is better off changing its behavior, even though a more fortunate state for everyone exists and is reachable.

Consider a 2-player symmetric game — a game where players share the same set of legal actions and reward function — where players can choose between cooperate (C) and defect (D). For every combination of actions $\{C_1C_2, C_1D_2, D_1C_2, D_1D_2\}$ there exists a corresponding reward that we shall name, from player 1‘s perspective and in order, reward (R), sucker (S), temptation (T) and punishment (P). In these settings, it is possible to define a social dilemma by adjusting the relative values of the rewards.

Let‘s take the most famous social dilemma as an example. The Prisoner‘s dilemma (PD) is a 2-player symmetric game, formally defined by the relative reward values such that the inequalities $T > R > P > S$ and $2R > S + T$ are satisfied. The background story accompanying this formal definition is a tale of two outlaws caught by the police and sent to different rooms for interrogation. Each prisoner has the option to snitch the partner (defect), for which he would receive a lesser punishment at the cost of a more severe punishment to his accomplice, or keep quiet (cooperate). Since the punishment for being snitched is greater than the penalty reduction for snitching, mutual cooperation is preferable to mutual defection.

If we use the values $R = 2$, $T = 3$, $S = 0$ and $P = 1$, it is straightforward to see how the dilemma plays out. Both players have the individual incentive to defect regardless of what the other will do, since the value they get for defecting is greater, either in case the other chooses to cooperate — $T = 3 > R = 2$ — or defect — $P = 1 > S = 0$. We call defect in the Prisoner‘s dilemma game a *dominating strategy*, that is, a strategy that yields greater reward than any other [Axelrod 1984]. In case both players choose the dominating strategy, they both get a reward of 1, even though there exists a more advantageous scenario for both players; one where they would have chosen to cooperate, and for that, would have gotten a reward of 2.

It is possible to create two other social dilemmas by changing the relative payoffs of R, T, S, and P [Kollock 1998]. If they satisfy the inequality $R > T > P > S$ we have

the game of *assurance* or *stag hunt*, while if they satisfy the inequality $T > R > S > P$ we have the game of *chicken*. Note that both games are still considered social dilemmas since they both have at least one deficient equilibrium.

3. Axelrod‘s tournaments

We call the one-time iteration of a Prisoner‘s dilemma a *one shot* PD, and there is not much to learn from this atomic interaction; as we have seen, the optimal strategy is to defect. The one-shot PD paints a bleak picture for our hopes of achieving mutual cooperation, but it certainly does not paint the whole picture.

In case the relationship between two players is extended, by incrementing the number of future encounters between them, the choice to defect cease to be trivial. We call the successive iterations of a PD game a *repeated* Prisoner‘s dilemma (rPD) [Gotts et al. 2003]. The difference between the one-shot and repeated games comes from the fact that, in the repeated case, any player can use previous interactions to guide future decisions. For instance, player 1 may punish player 2 in a future round for not cooperating in this current round. This ”shadow of the future” effect [Axelrod 1984] changes the dilemma in its core; the decision to C or D has to be made considering the effect it might have on the opponent‘s decision in future iterations.

The natural question that emerges is: what is the dominating strategy to guide one‘s decision in such a scenario? There is no dominating strategy that is independent of the opponent‘s strategy [Axelrod 1984]. We can quickly build a counterexample that works as proof: Suppose we play against an opponent that always defects (ALL-D). The best strategy we can use is to mirror our opponent‘s strategy and also always defect. Now, suppose we play against a player that always cooperates until defected against. In this case, our best option is to cooperate for at least a while before we think about defecting — if we think about defecting at all.

However, to say that there is no best strategy independent of the opponent‘s strategy does not mean that all strategies are equally good. In a seminal body of work, political scientist Robert Axelrod promoted two tournaments where game theorists could send strategies to play against one another [Axelrod 1980a] [Axelrod 1980b]. The study was an attempt to shed light on the question of how to properly behave in an rPD so as to maximize individual return [Axelrod 1980a].

The first tournament was run in a round-robin format, such that each of the fourteen entries would play against one another, against a RANDOM strategy, and against itself. Each game consisted of 200 iterations. Surprisingly, the winner of that tournament was the simplest of all strategies submitted, called TIT-FOR-TAT (TFT), submitted by professor Anatol Rapoport [Axelrod 1980a]. TFT cooperates in the first round and copies the action taken by the opponent in the previous round henceforth. Phrasing it differently, TFT is a purely responsive strategy, and it matches cooperation with cooperation and defection with defection, the response delayed by one round.

The second tournament had a similar format to the first, with the exception that the game had a small probability of ending in each given move, so strategies couldn’t exploit the fact that they knew when the game was going to end and throw sneaky end-game defections. This tournament had 62 entries from 6 different countries, ranging from a 10-year-old computer hobbyist to professors of computer science, economics, psychology,

mathematics, political science, and evolutionary biology [Axelrod 1980b]. Once again, TFT claimed first place.

3.1. Key learning points

As one can imagine, it was not by fate or luck that TFT did so well in both tournaments. Axelrod points out four properties shared by successful strategies, including TFT [Axelrod 1984]:

- *Niceness:* Successful strategies are never the first to defect. Nice strategies do especially well against other nice strategies since the threat of exploitation does not exist.
- *Forgiveness:* Successful strategies can forgive the opponent and cooperate after being defected against. Forgiveness is of interest to the forgiving part since not forgiving might mean missing out on future mutual cooperation.
- *Provocable:* Successful strategies are provocable, that is, they can retaliate against an opponent‘s act of defection immediately after the fact. Retaliation may signal to others low tolerance towards defection.
- *Clear:* Successful strategies are clear and simple to understand.

These four properties present a general summary of what is needed for an all-around strategy. Note that for a strategy to be successful, it needs to perform well against different types of strategies, including those that are exploitative, overly-forgiving, overly-provocable, etc. TFT has all four of these properties and proved to be robust in all match-ups it encountered [Axelrod 1980b].

4. N-player social dilemmas: public goods and the tragedy of commons

The 2-player social dilemma framework is a good primer to the problem of conflicting individual and collective incentives. However, it is not sufficient to represent the whole range of dilemmas we face in the real world, especially those at the communal level. A natural extension to it is the n-player social dilemma framework, most often represented by two metaphorical stories of ”mythic” proportions [Kollock 1998]: Public Goods and The Tragedy of the Commons. Similar to the 2-player case, both games may be represented as a reward tensor that maps every combination of cooperation and defection to a reward for each participant.

4.1. Public goods

The public goods dilemma is named after a set of problems where a group of individuals needs to pay an upfront cost to utilize a public resource. Once again, every individual has the incentive not to pay the fee and free-ride, though, if everyone does so, there will be no public good to enjoy from [Kollock 1998].

A relatable example of a public good dilemma comes from the municipal tax system. The local government employs the money from our taxes, among other expenditures, to improve public spaces that every resident has the right to use. The lack of payment from one resident will most likely not affect the maintenance of our roads and parks, but if tax evasion becomes a norm, the community — especially those who paid their taxes — will be punished by having public spaces that are not well maintained [2].

[2] We are simplifying the problem to make our point and not accounting for the fact that if you don‘t pay your taxes, you are likely to be punished, which in the real world changes the payoff matrix.

4.2. Tragedy of the commons

The tragedy of the commons is a term introduced by Garret Hardin [Hardin 1968] to describe a set of social problems that cannot be solved solely by technological advances; instead, a behavior change is needed. Like the public goods dilemma, the tragedy of the commons is a group dilemma. Still, unlike the former, it is associated with individuals being incentivized to increase own short-term payoff at the cost of inflicting a punishment to everyone in the group.

Hardin himself describes a didactic example in his original article. A group of herders, having access to a common piece of land, can allow as many of their cows graze on it. Every herder has the individual incentive to let as many cows in, but if all herders behave accordingly, the grass will soon be depleted, and the cows will have nothing to eat [Hardin 1968].

4.3. Theoretical predictions and empirical evidence of n-player dilemmas

The theoretical findings regarding the group dilemmas described in this section point to a non-endogenous resolution to this set of problems [Hardin 1968, Ostrom 2000]. The theoretical framework is based on three assumptions *a)* Resource users are norm-free utility maximizers with no bounded rationality; *b)* Designing rules to change incentives is an easy task; *c)* The resolution to these problems demands intervention from a central authority [Ostrom 1999].

Nonetheless, empirical work has provided evidence that these three assumptions don‘t conform with reality. In practice, many experimental studies have shown instances of n-player dilemmas being solved by local communities, without the need of a regulatory central authority, and that social norms play a significant role in solving them, as shown by Economics Nobel Prize winner Elinor Ostrom [Ostrom 1999, Ostrom 2000]. The theoretical endeavor that tries to bridge the gap between the expected defection of participants and empirical evidence taps from evolutionary theories and points towards a complex system framework, with agents of multiple personalities interacting and being able to adjust behavior based on outcome [Ostrom 1999, Ostrom 2000].

5. Agent-based simulations of social dilemmas

The 2-player and n-player social dilemmas presented so far are well structured — they are a simple and deterministic mapping from actions to rewards — which makes them good candidates for agent-based modeling. This section presents a small sample of studies that made use of agent-based simulation to examine some properties of social dilemmas.

Deadman used the multiagent framework Swarm to simulate adaptive agents with limited rationality in a common-pool resource experiment [Deadman 1999]. Agents in the simulation were conceived as having two components: a model of the environment that consisted of state transitions given past actions, and a model of themselves, that contained instructions for generating behavior. Significant variance in strategies‘ performance was noted as a function of agents‘ recent actions, and no single strategy emerged as dominant.

Izquierdo et al. used a linear stochastic model of reinforcement learning to study the dynamics of two-player, stochastic-strategy social dilemmas [Izquierdo et al. 2008]. In each round, players had to choose whether to cooperate or defect and revise their current probabilities based on a threshold value called *aspiration level*. The revision would

be upwards if the payoff stayed above the aspiration level and downwards otherwise. They report two types of attractors within these systems; a self-reinforcing equilibrium, when agents enter a virtuous cycle of mutual cooperation, and a self-correcting equilibrium, loosely defined as a stable equilibrium in the vicinity, i.e., trajectories that get sufficiently close will be drawn to it as time approaches infinity.

Guerberoff et al. studied the effects of language representation expressiveness in spatial rPD simulations [Guerberoff et al. 2011]. Cooperation between agents increased as expressiveness grew in complexity. Queiroz and Sichman used parallel computing to extended this work and tested the effects of environment variables such as grid size, mutation rate, and error rate [Queiroz and Sichman 2016].

Barbosa et al. employed RL techniques and agents with varying cognitive capabilities — defined as the size of the state space over which strategies can be learned — in an n-player Prisoner‘s dilemma game [Barbosa et al. 2020]. They tested the effects of environment variables, learning variables, and agents‘ cognition on the outcome of an n-person Prisoner‘s dilemma simulation. We highlight the results on the impact of cognition on cooperation rates. Surprisingly, there is no positive correlation between an increase in cognition, as defined, and the rate of cooperation after the cognitive capacity surpasses a small threshold.

6. Modelling social dilemmas as Markov Games

We refer to the matrix formalization of social dilemmas presented thus far as the high-level model (HLM). The justification behind this label comes from the series of abstractions needed to fit real-world problems into this formalism. Leibo et al. summarizes some key differences: *a)* In the real world, the set of actions is not a clear dichotomy made of cooperation and defection. Cooperation and defection are rather implicit in one‘s actions and may occur in a continuum. *b)* In the real world, actions are not perfectly synchronous, and minimal asynchronies may signal one‘s intentions to others, which may, in turn, change their behavior. *c)* The world is a complex place, and hardly ever all information needed to guide one‘s decision is available [Leibo et al. 2017].

More recently, a model that circumvents the issues above has been proposed to tighten the gap between HLMs and the real world. Hereafter, we refer to it as the low-level model (LLM), since it captures with greater detail the process of a human dealing with a real-world social dilemma. The model is based on the Markov game framework, which is an extension of the Markov Decision Process (MDP) for the multiagent setting [Littman 1994].

A Markov game is generally defined by a set of states S, a collection of action sets $A_1, A_2, ...A_n$ one for each agent, a transition function $T : S$ x A_1 x A_2, ..., $A_n \longrightarrow$ PD(S), that maps combinations of states and actions to a probability distribution over S, and a reward function $R : S$ x A_1 x A_2, ..., $A_n \longrightarrow \mathbb{R}^n$ that yields an immediate reward to each agent. Each agent acts so as to maximize its expected long term reward, discounted by a factor of γ at each time step, that is, $\mathbf{E}\left[\sum_{j=0}^{\infty} \gamma^j r_{t+j}\right]$.

From now on, we turn our attention to a series of studies done using the LLM to characterize the dilemmas. The work reviewed in this section present how mixed-motive games are being simulated with the advent of RL algorithms and reveal a new and exciting

avenue of research. They are mainly divided into two categories: *a)* studies that try to test some hypothesis about a social dilemma by exploring the dynamics of the system, and *b)* studies that propose a new agent architecture and most often try to induce reciprocity to achieve mutual cooperation. Before we start the review, we introduce the environments used in the simulations as we believe it will facilitate the main ideas present in a LLM.

6.1. Environments

The first environment we describe, named *commons* [Pérolat et al. 2017], mimics a real-world tragedy of the commons, and consists of a 2D grid world where agents harvest apples, represented by colored cells, to earn rewards. Each cell has an apple spawn rate proportional to the number of apples left in its vicinity, so the fewer apples, the lower the spawn rate. Agents within this environment choose actions such as: moving up, down, left, right, rotating left and right, and collect apples by stepping into cells that contain one. They can also inflict a cost to other agents by shooting a laser beam that tags others out of the game for a predefined number of time steps. In order to maximize own reward, agents need to manage their greed so as not to deplete the environment completely.

The second environment used in simulations, named *coins* [Lerer and Peysakhovich 2018], mimics an n-player Prisoner‘s dilemma. The environment consists of a 2D grid world where agents have to collect coins that spawn in random cells at a fixed rate to receive an immediate reward. Agents and coins in the environment have a color property, and each agent receives a reward of +1 every time it collects a coin and a punishment of -1 every time another agent collects a coin that shares its color property. Actions are agent-centered and very similar to the commons game. Cooperation in the coins game means only collecting coins of the same color as its own as this behavior maximizes total utility.

The third environment, named *wolfpack* [Leibo et al. 2017], resembles the stag hunt game briefly described in section 1.1. The game once again occurs within a 2D grid world, and the players‘ action set is similar to the previous games. The n-players are divided into two groups: the wolves and the prey, and the game unrolls as the wolves try to catch the prey. Once the prey is caught, rewards are distributed to the wolves proportional to the number of wolves in the region the catch took place. This game is different in its payoff structure than the others; here, the reward received for mutual cooperation is greater than the one obtained for unilateral defection.

Finally, the *clean-up* environment [Leibo et al. 2017] mimics the public goods game, and similar to the commons game, agents earn rewards by harvesting apples. The apples‘ spawn rate in this environment is inversely proportional to the pollution of a nearby river and drops to zero once a threshold value is surpassed. Pollution increases with fixed probability over time, demanding agents to clean it up every now and then so apples keep spawning. Every time an agent cleans the river, it spends time it could otherwise be harvesting apples, thus the dilemma.

6.2. Dynamics of the system

Leibo et al. proposed a framework entitled Sequential Social Dilemmas, based on empirical game-theoretic analysis, that is built on top of the Markov game framework [Leibo et al. 2017]. The framework considers cooperation and defection as policy properties instead of atomic actions and preserves the payoff structure seen in the matrix-based

formalism for each game. A policy is deemed to be cooperative or defective based on a *social behavior metric* value, which is independent of the state value function. The study used the framework to analyze the learning dynamics of the system in two environments: wolfpack and commons. In the commons game, the social behavior metric used was the frequency of laser beams shot. Results pointed to an escalation in conflict as the resources became scarcer and conflict costs increased. The experiment on the wolfpack environment used wolves‘ closeness as its social metric. Two different cooperative policies emerged from the simulations: one where wolves looked for each other before hunting the prey and the other, where a wolf would first find the prey and then wait for the other wolf to arrive.

Perolat et al. also tested the learning dynamics of simulations in the commons environment and analyzed the emergent social outcomes in each learning stage [Pérolat et al. 2017]. Agents in the simulation learned through self-play via a Q-learning algorithm with function approximation. The social outcomes considered were: utility (U), sustainability (S), equality (E) and peace (P), and measured the aggregate effects of individual actions. The authors describe three distinct learning phases: *naïvety*, *tragedy* and *maturity*. Phase 1, *naïvety*, occurs at the beginning of training and is characterized by high commons‘ stocks. Due to agents‘ lack of efficiency at harvesting apples, stocks are never depleted, and U and S are relatively high. P is also high since agents see no advantage in tagging others. Phase 2, *tragedy*, is characterized by the rapid depletion of stocks. Agents in this phase become too good at harvesting apples, and because of that, U and S sharply fall. Finally, phase 3, *maturity*, is characterized by an increase in conflict and a decrease in P. Agents at this point realize tagging can be useful in scenarios of scarcity, and because of that, the other three social outcomes increase.

Starting from the premise that many humans have inequity-averse preferences, Hughes et al. trained agents with these social preferences on the commons and clean-up environments [Hughes et al. 2018]. Both advantageous and disadvantageous inequity aversion were considered, which are, respectively, reductionist models of guilt and envy. This social value is modeled as a reward layer on top of each environment‘s extrinsic payoff structure and as the name suggests, penalizes agents as the difference between their rewards grows further apart. Results show the differences in behavior between advantageous and disadvantageous inequity averse agents. Advantageous inequity aversion drives the difference in behavior from the self, whereas disadvantageous inequity aversion induces mutual cooperation by punishing defection. Furthermore, it was found that advantageous inequity aversion is particularly effective for resolving public goods dilemmas, whereas disadvantageous inequity aversion performs better in the commons dilemma.

Finally, McKee et al. studied the effects of social diversity in mixed-motive games [McKee et al. 2020]. Similar to the work of Hughes et al., they embedded social preferences in agents as an intrinsic reward layer on top of the extrinsic rewards coming from the environment. Preferences were modeled following the Social Value Orientation (SVO) framework from interdependence theory. Simulations were conducted in the commons and clean-up environment. As expected, homogeneous altruistic populations of agents performed well compared to other homogeneous populations in both the commons and clean-up games but scored surprisingly low on the equality metric. Populations with diverse SVO, on the other hand, performed well in both cases while maintaining a higher

equality score.

6.3. Novel architectures

Lerer and Peysakhovich proposed an agent‘s architecture named amTFT that emulates the behavior of TFT in a two-player, extended Markov game setting, the coins environment [Lerer and Peysakhovich 2018]. The amTFT architecture comprises two policies: a purely cooperative and a purely defective. Cooperative policies were defined as those that, starting at an initial state s, maximize the joint expected reward $V^1(s, \pi_1, \pi_2) + V^2(s, \pi_1, \pi_2)$. The cooperative and defective policies were trained by means of self-play in two reward schedules: a purely cooperative and a purely selfish. An amTFT agent works by switching from cooperation to defection for k time steps — k being a function of the agent‘s tolerance towards defection — every time it detects defection by its counterpart. Cooperation/defection is measured at each time step in terms of the partner‘s relative Q-values; the Q-value given the partner‘s action versus the Q-value of the action the partner would have taken if it had cooperated.

Eccles et al. also proposed an agent‘s architecture that implements reciprocity in Markov Games [Eccles et al. 2019]. Three points of improvement were listed compared to Lerer and Peysakhovich‘s work: *a)* reciprocity is not defined only for fully cooperative or fully defective actions, it can be defined in the cooperation-defection continuum; *b)* reciprocity is not only defined for the 2-player case; *c)* reciprocity should not depend on observing the rewards of others; *d)* reciprocity can be learned online. The work accomplishes the proposed improvements by implementing two types of agents: innovators and imitators. An innovator optimizes for purely selfish rewards while an imitator measures and matches the sociality level of others. This diversity creates an interesting dynamic; innovators are incentivized to cooperate since their sociality will be matched by imitators. Sociality, as proposed, is measured in terms of a niceness function that maps the effects of one player‘s actions onto others, similar to the notion of a value function.

7. Discussion

7.1. Differences between models and the issue with cooperation

LLMs and HLMs differ in complexity, the former having a greater grasp of real-world dilemmas than the latter. It is important to state two key differences between them that bring the LLM formalism closer to reality: *a)* HLM is stateless and represented by a simple mapping from actions to rewards, while LLM encompasses state transitions. In practice, it means the rewards in LLM are dependent on the agent‘s sequence of state-action pairs. *b)* There exists a one-to-one mapping between actions and the notion of cooperation. That is to say that agents within HLM opt explicitly to act more or less cooperative, while in LLM cooperation or defection are implicit in one‘s sequence of actions. This has implications to cooperation‘s continuity and synchrony (see 1st. paragraph, items *a* and *b* of Section 6).

An arguably deeper problem emerges when switching from HLMs to LLMs, and it has to do with cooperation changing from being an explicit action to an implicit meta-action. How does one measure cooperativeness? Leibo et al. [Leibo et al. 2017] and Perolat et al. [Pérolat et al. 2017] measured cooperation through a domain-dependent social behavior metric. The problem with this approach is that it is bias-inducing prone

and is only generalizable to problems with a social metric that can change the game's payoff structure as it varies. Furthermore, using a social behavior metric to measure cooperation may lack meaning. For instance, both authors in [Leibo et al. 2017] and [Pérolat et al. 2017] used the frequency of tagging as a proxy for cooperativeness in the commons environment; should an agent that does not tag others but completely depletes the environment be considered truly cooperative?

Conversely, Lerer and Peysakhovich measured cooperativeness based on expected outcome [Lerer and Peysakhovich 2018]. This approach fails to generalize to the n-player case as pinpointing defectors becomes impossible. Furthermore, in stochastic processes — such as many real-world dilemmas —, we can think of scenarios where cooperativeness may yield bad results.

In Eccles et al., the authors use a sociality function that better approximates the meaning of cooperation in the real world [Eccles et al. 2019]. Still, this approach is not generally agreed upon, i.e., a theory of cooperation is yet to be fully developed.

7.2. How to better reciprocate?

As we have seen throughout this paper, reciprocity, embodied as TFT in the early days of Axelrod's tournaments, provides a suitable mechanism for solving social dilemmas. Other attempts at simulating reciprocal behavior, such as the Tit for 2 Tats strategy did not achieve the same level of robustness [Axelrod 1980b]. But the question of how to best reciprocate is still open, and experimenting with Markov games may provide us an answer. It is natural to stick to what has worked in the past, but could strategies such as Tit for 0.9 Tats or Tit for 1.1 Tats perform better in certain situations? These scenarios are open to investigation.

8. Conclusion

The importance of cooperating in mixed-motive settings is highlighted by [Dafoe et al. 2020]. The authors discourse about future challenges we will face as a society, all orbiting around the ill-defined notion of cooperation in mixed-motive environments and which they termed Cooperative AI. To illustrate with an example, how will an autonomous vehicle interpret the fast-walking pedestrian crossing the street looking at his/her phone? Cooperating with this pedestrian might mean something different than cooperating with the honking driver behind who is late for work. As we can see, many of these issues are immersed in social norms, and solving them is far from trivial.

Advancing this newly conceived field is of great concern for many areas of science. In the 80s and 90s, analyzing how artificial agents interacted in these situations could provide insights into acting in real-world societies. As of today, besides guiding us in some of our social interactions, solving these issues could represent a significant gain in productivity as the population of artificial agents acting on behalf of humans grows in size.

9. Acknowledgements

This work was carried out with the support of Itaú Unibanco S.A., through the scholarship program Programa de Bolsas Itaú (PBI), linked to the Centro de Ciência de Dados (C^2D) of Escola Politecnica da USP.

References

Axelrod, R. (1980a). Effective Choice in the Prisoner‘s Dilemma. *Journal of Conflict Resolution*, 24(1):3–25.

Axelrod, R. (1980b). More Effective Choice in the Prisoner‘s Dilemma. *Journal of Conflict Resolution*, 24(3):379–403.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic, New York.

Barbosa, J. V., Costa, A. H. R., Melo, F. S., Sichman, J. S., and Santos, F. C. (2020). Emergence of Cooperation in N-Person Dilemmas through Actor-Critic Reinforcement Learning. *Adaptive Learning Workshop (ALA), AAMAS 2020 - Autonomous Agents and Multi-Agent Systems*.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. (2020). Open problems in cooperative ai.

Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology*, 31(1):169–193.

Deadman, P. J. (1999). Modelling individual behaviour and group performance in an intelligent agent-based simulation of the tragedy of the commons. *Journal of Environmental Management*, 56:159–172.

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., and Leibo, J. Z. (2019). Learning reciprocity in complex sequential social dilemmas.

Gotts, N. M., Polhill, J. G., and Law, A. N. (2003). Agent-based simulation in the study of social dilemmas. *Artificial Intelligence Review*, 19:3–92.

Guerberoff, I., Queiroz, D., and Sichman, J. (2011). Studies on the effect of the expressiveness of two strategy representation languages for the iterated n-player prisoner's dilemma. *Revue d'Intelligence Artificielle*, 25:69–82.

Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859):1243–1248.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., and Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Izquierdo, S., Izquierdo, L., and Gotts, N. (2008). Reinforcement learning dynamics in social dilemmas. *Journal of Artificial Societies and Social Simulation*, 11.

Kollock, P. (1998). Social dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, 24(1):183–214.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, page 464–473, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Lerer, A. and Peysakhovich, A. (2018). Maintaining cooperation in complex social dilemmas using deep reinforcement learning.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94, page 157–163, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

McKee, K. R., Gemp, I., McWilliams, B., Duèñez Guzmán, E. A., Hughes, E., and Leibo, J. Z. (2020). Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 869–877, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Ostrom, E. (1999). Coping with tragedies of the commons. *Annual Review of Political Science*, 2(1):493–535.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3):137–158.

Pérolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Queiroz, D. and Sichman, J. (2016). Parallel simulations of the iterated n-player prisoner's dilemma. In Gaudou, B. and Sichman, J. S., editors, *Multi-Agent Based Simulation XVI*, pages 87–105, Cham. Springer International Publishing.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.