Reinforcement Learning in Conflicting Environments for Autonomous Vehicles

Dominik Meyer, Johannes Feldmaier and Hao Shen

Abstract—In this work, we investigate the application of Reinforcement Learning to two well known decision dilemmas, namely Newcomb's Problem and Prisoner's Dilemma. These problems are exemplary for dilemmas that autonomous agents are faced with when interacting with humans. Furthermore, we argue that a Newcomb-like formulation is more adequate in the human-machine interaction case and demonstrate empirically that the unmodified Reinforcement Learning algorithms end up with the well known maximum expected utility solution.

1 Motivation

Autonomous Unmanned Underwater Vehicles (UUVs) are used for a wide range of oceanographic, maritime mining, and military tasks including underwater surveys, inspection and maintenance of submerged structures, tracking oceanographic features, and undersea mapping to name a few (cf. [8]). Depending on their task, the physical shape of UUVs can be the traditional torpedo-shaped bodies or more like Remotely Operated Vehicles (ROVs) with many manipulators, cameras, and lights.

The difficulty with autonomous submarines is that there is no communication link between a human operator due to the fact that radio waves cannot penetrate water very far. Therefore, the UUV loses its GPS signal as soon as it enters the water. Navigation is mostly performed using compasses, depth sensors, accelerometers, and sonars. The sensor information provides enough information for navigation using dead reckoning but are not sufficient to make an informed decision.

In the upcoming deep sea mining scenarios, where wellheads are built as subsea systems directly on the sea floor, ROVs and UUVs are used to build and maintain those structures. As direct human control is only possible with tethered ROVs, an increased usage of UUVs to maintain subsea systems is desirable. Using underwater acoustic positioning systems (long-baseline (LBL) systems), UUVs are able to find subsea structures like wellheads and processing systems. Only acoustic communication allows limited communication like broadcasting alarms.

In such a scenario an unmanned underwater vehicle needs advanced autonomous decision making algorithms for task planning and behavior selection. Furthermore, in case of multiple cooperating robots and operator controlled vehicles, strategies for cooperating and solo actions have to be developed and risk analysis is vital [10]. Besides traditional risk assessment using event and fault tree analysis, risk can also be assessed using examples from game theory like the Prisoner's Dilemma or the Newcomb's Problem. We investigate these examples by applying Reinforcement Learning (RL).

Reinforcement Learning, one classic decision making and learning technique in the field of autonomous learning, fits the setting of the two dilemmas very well, and has been already applied in basic Prisoner's Dilemma scenarios. In [2], the authors apply RL to multi-player domains, where cooperation is beneficial and investigate the capability of two agents successfully establishing a stable equilibrium strategy. The authors of [1] present the performance of RL algorithms in the Prisoner's Dilemma, admitting knowledge of the problem structure. They distinguish between independent learners, where each agent has no knowledge about the state of the other agents, and multi-agent settings, where a joint decision is reached. They aim at establishing cooperation, which can be reached with a biased exploration strategy. Sandholm et al. [7] study the play of a RL agent against a fixed opponent strategy and itself. They manage to achieve optimal play by using a history of moves and a representation of the move history by a neural network as the state. On the contrary, Flache et al. [4] take a different approach, which tries to explain the cooperative behavior observed in experiments with a general psychologically inspired reinforcement learning model.

2 Modelling Deep Sea Robots using Newcomb-like Problems

Newcomb's problems arise when an autonomous agent is in a situation where others have knowledge about its decision process via some mechanism (e.g. a statistical based model) that is not under its direct control. Newcomb-like problems cannot be handled by the conventional Causal Decision Theory, as the independence of the two decision makers is violated. Also most real decisions humans face are Newcomb-like, at least whenever other humans are involved. People automatically involve their experience and read unconscious or unintentional signals in order to build an internal social model of how someone decides. Simultaneously, they use those models to make their own choices and this when Causal Decision Theory fails.

• *The authors are with the Chair for Data Processing, Department of Electrical and Computer Engineering, Technische Universität München. E-mail: <firstname.lastname>@tum.de.*

Manuscript submitted June 13, 2016.

In general, real world decision scenarios can be often described as Newcomb-like problems in several ways. We however, do not assume a priori knowledge of the problem structure and instead treat the decisions made as a variation of a two-armed bandit. Furthermore, we restrict ourselves to the setting of independent learners. Namely, two actors, which can be characterized by their probability of cooperation in the respective scenario. This is possible, as we can establish the equivalence of this specific version of the Prisoner's Dilemma with Newcomb's Problem.

2.1 Prisoner's Dilemma in Deep Sea Repair Robots

In our example of a field of multiple oil wells and two available maintenance robots, two failures with smaller leakages occur at the same time. Each leakage can be fixed by one robot, but there is a chance of the robot to be damaged in the process. Also, if the leakages are both not fixed, there will be a cascade of events that will cause an even bigger oil spill. If both robots fix both leaks, then there will be no oil spill but still the chance of slight damage to the robot remains. If one robot tries to fix one spill and the other decides not to, then there will be no natural desaster, but the second leak will still remain. Also the robot trying to fix the spill will be severely damaged as the second spill causes complications with the first leak. Underwater, the robots have no communication to coordinate whether they will go for a repair or not, but can sense the location of the leaks and the other robots with a sonar.

An exemplary payout matrix (regrets are treated as negative payouts) for this problem is depicted in Table 1, with $R = -2.000$, the cost of the repair and eventual minor damages of the robot, $S = -4.000$, the cost of major damages to the robot due to the overpressure from the other leak as well as the costs for the oil spilled, $T = -1.000$, the costs of the oil spill by not fixing the leak, $P = -3.000$, the cost of the major oil spill due to both leaks not getting fixed. Clearly, it satisfies all the prereqisites to be a version of the Prisoner's Dilemma with having the two conditions

$$T > R > P > S \qquad \text{and} \qquad R > \frac{T+S}{2}. \tag{1}$$

fulfilled. The dilemma unfolds as follows: since the temptation T is the lowest regret a robot can receive while the sucker's regret S is the highest, the optimal strategy for a robot locally would be to always not repair (defect) since the action of the other robot is unknown. The robot therefore has no incentive to change its decision, if it is unclear what the other will do. This situation is also known as Nash equilibrium in game theory.

TABLE 1: Oil Spill Prisoner's Dilemma Robot-Robot Regrets.

		Robot 2	
		Repair	No Repair
Robot 1	Repair	(R, R)	(S, T)
	No Repair	(T, S)	(P, P)

If both robots would know what the other does or would be able to communicate about a common strategy, then clearly mutual cooperation (both repair) would be the optimal thing to do. Herein lies the dilemma. Locally, it is optimal for each robot to not repair while globally it would be optimal to repair each leak. For this dilemma to occur, we need the first condition to be fulfilled. Then the second condition prevents taking turns at defection and cooperation to be more profitable in an iterated setting.

A robot could now concretely decide what to do using the expected utility of each outcome. This means that it assigns a certain cooperation probability p to the problem, which gives the likelihood that the other robot will choose `repair`. As the problem is symmetrical, we can write down the expected utility EU as

$$\begin{cases} EU(\text{repair}) = p * (-2.000) + (1-p) * (-4.000), \\ EU(\text{no repair}) = (1-p) * (-1.000) + p * (-3.000). \end{cases} \tag{2}$$

If therefore the chance of the other robot cooperating is greater than $p = 0.75$ then the robots will cooperate and both choose `repair`.

2.2 Imbalance in Prisoner's Dilemma due to Human Interaction

If now a human happens to be involved in the dilemma, the situation changes drastically. Suppose a maintenance worker is underway in a single person submarine to carry out maintenance tasks that cannot be handled by autonomous robots yet. In this moment the situation described prior unfolds and two leaks have to be fixed. Compared to the regret, when a robot gets damaged, the regret, when the submarine carrying the human gets damaged, is much higher. The event in which the human solely decides to fix one leak and an accident happens is now valued with a regret of $-1.000.000$. The modified payout matrix is written down in Table 2. As can easily be seen, the regret situation is now highly skewed, since the life of a human naturally is weighted much higher than the integrity of the replaceable robot. Therefore, the robot would still decide to repair according to expected utility starting from a confidence of $p_r = 0.75$ that the human will also repair, whereas the human would only decide to go down for a repair if she/he is almost sure with a confidence of $p_h = 0.999$ that the robot will follow and do it's job.

TABLE 2: Oil Spill Prisoner's Dilemma Robot-Human Regrets.

		Human	
		Repair	No Repair
Robot 1	Repair	$(-2.000, -2.000)$	$(-4.000, -1.000)$
	No Repair	$(-1.000, -1.000.000)$	$(-3.000, -3.000)$

2.3 Modelling a Skewed Prisoner's Dilemma with Newcomb's Problem

The classical Prisoner's Dilemma cannot fully accompany skewed problems of this type. But fortunately, as Lewis stated [5], the Prisoner's Dilemma can be seen as two coupled Newcomb's Problems. In the original description of the Newcomb's Problem, a superhuman intelligence is predicting the choice the player is going to make and placing the bets on the two options available accordingly. The coupled Newcomb's Problem from the viewpoint of the Robot in the

TABLE 3: Oil Spill Dilemma Newcomb Version, Robot Viewpoint.

	Robot	
	Repair	No Repair
Predict: Human Repair	-2.000	-1.000
Predict: Human No Repair	-4.000	-3.000

asymmetrical dilemma can be observed in Table 3. In our version, instead of the superhuman intelligence predicting our choices, we predict how the opponent (in this case the human) in the Prisoner's Dilemma version of the problem will behave. Depending on the prediction accuracy, the payout can be observed upon choosing one of the actions. Originally, the prediction probability was set very close to $p = 1.0$ and therefore the dilemma again manifests itself in that the expected utility recommends to do the opposite of the dominating choice. By following the argumentation in the work of Nozick [6], the robot now can choose to not repair, since both outcomes of not repairing are better than their alternatives, no matter if the prediction was correct or not. If we insert the numbers in the above formulas for expected utility then, on the other hand, with a very high probability that the prediction was correct, it is better to choose to repair.

Using this modified problem formulation, we can now write down the regrets from the viewpoint of the human in Table 4. As these are now two decoupled problems, it is additionally possible to assign different cooperation probabilities to the two involved entities. Usually, we would assign a very high cooperation probability to a robot as it should do what it is supposed to do, but in the setting of autonomous decision making with limited communication under water, we cannot be so sure anymore. Also for humans, depending on the situation, cooperation probability can exhibit a great variance.

TABLE 4: Oil Spill Dilemma Newcomb Version, Human Viewpoint.

	Human	
	Repair	No Repair
Predict: Robot Repair	-2.000	-1.000
Predict: Robot No Repair	$-1.000.000$	-3.000

3 Simulation Results using Reinforcement Learning

To assess how an autonomous robot would learn to decide in a deep sea repair scenario, we can now use the above models to simulate behavior. The autonomous learning technique we chose to test is unmodified Reinforcement Learning, to get a baseline of the most popular autonomous learning algorithms that currently exist in the literature.

It is important to notice that in our setting of decision making, one decision maker plays against her/his opponent with a fixed cooperation or prediction probability. In other words, there is no state information required for both Newcomb's or the Prisoner's Dilemma type of formulation. Hence, we model our problems as playing a bandit. Specifically, for each action $a \in A_n := \{\texttt{repair}, \texttt{ no repair}\}$ we maintain the Q function without the state variable as $Q(a) = \mathbb{E}\left[\sum_t \gamma^t r_t\right]$, which represents the future expected discounted reward, where t is the current iteration step, $\gamma \in [0, 1]$ a discounting factor and r_t the payout received at iteration step t. The two problems are evaluated in an iterated fashion. This means that the agent is not faced with the decision only once, but for N times and can therefore learn from those interactions.

For updating the Q values, we use the well known SARSA [9] algorithm with learning rate α. A second learning agent is calculating averaged payouts for the Q function (AVGQ). A third agent calculates expected utility (EU) for comparison. As the action selection mechanism, we chose an ϵ-greedy strategy, where the action with the largest corresponding Q value is chosen most of the times and a random action with probability ϵ. In each problem, learning was studied for $N = 10000$ iteration steps and averaged over 50 independent runs.

For the Newcomb formulation, we additionally implemented two baseline agents, that either always repair or don't repair. If now the prediction accuracy is varied, the payout behaves as expected, which can be seen in Figure 1a.

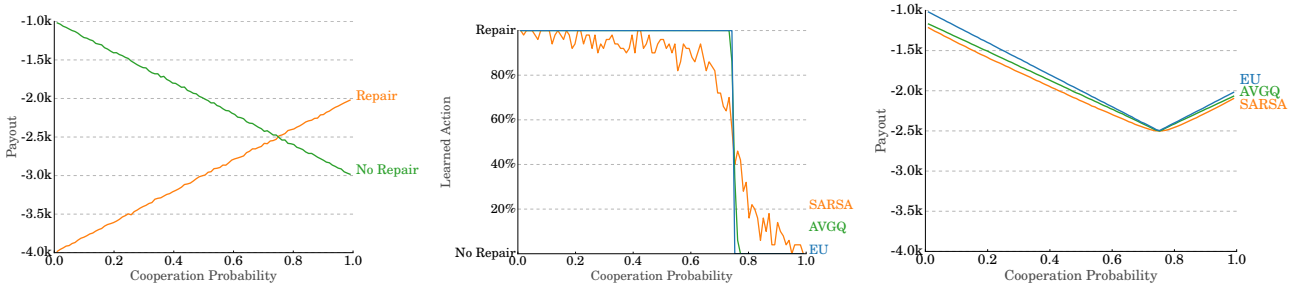
For the SARSA agent, we can observe adaptation to the problem depending on the prediction accuracy. In Figure 1b, we can observe that the agent will always choose `no repair` if the cooperation probability is close to zero, and always `repair` if the probability is very high. If this probability accuracy is approximately 0.75, then the agent cannot learn what to do and chooses actions at random. The RL agent therefore is able to learn the correct behavior depending on the environmental parameters. The average Q (AVGQ) implementation learns to behave in the same way, even closer to the expected utility solution. In accordance to this behavior, the payout varies as shown in Figure 1c.

What we can conclude from Figures 1b and 1c is that the RL agent is able to learn the action that maximizes the payout and corresponds with the expected utility solution. This means in a consistent environment where the partner almost always cooperates or does the opposite, learning will succeed. For cooperation probabilities around the expected utility threshold ($p = 0.75$), the behavior is not perceived as consistent and the best the SARSA agent can do is to choose actions in a random manner.

In the Prisoner's Dilemma formulation, we can imagine two modes of operation. Namely, either each agent receives only its own regret, or both agents receive the sum of the two regrets. In Figures 2a and 2b, it can be seen that for individual payouts `no repair` always dominates `repair`, while for the sum of payouts `repair` dominates. This reflects the dilemma in the underlying problem and it would therefore be optimal for the RL agents to learn either to not repair (in the individual payout setting) or to repair (in the sum-of-regret setting). This is in fact the case if we consider the results from Figure 2c, in which the regrets for the SARSA and AVGQ agents are depicted for the individual (I) and sum (T) regret setting.

4 Conclusion

In this abstract, we study the behavior of basic unmodified reinforcement learning agents, when faced with decision theoretic thought experiments. In both problems,

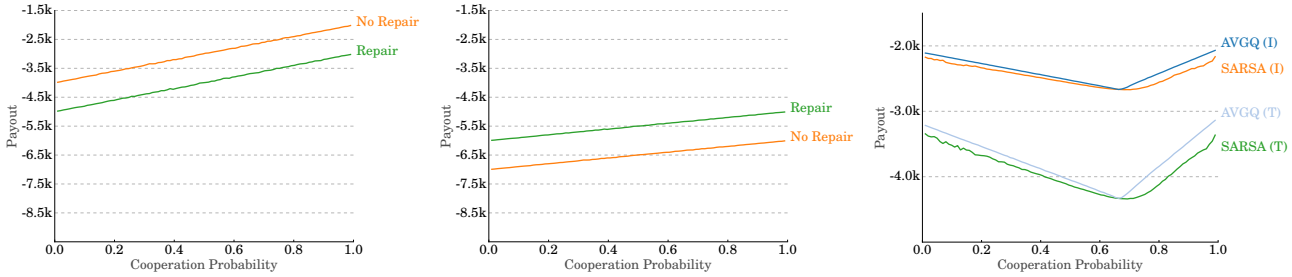


(a) Payout for two baseline agents that always repair or don't repair.

(b) Learned action for SARSA with $\alpha = 0.1, \gamma = 0.9, \epsilon = 0.1$, AVGQ and EU. Probability of repairing on y-axis.

(c) Payout for SARSA (same parameters), AVGQ and EU for varying cooperation probability.

Fig. 1: Experimental results for Newcomb's Problem, Robot View.



(a) Payout for two baseline agents that always repair/don't repair.

(b) Sum of payouts for the baseline agents.

(c) Payout for SARSA with $\alpha = 0.1, \gamma = 0.9, \epsilon = 0.1$ and AVGQ. (I) individual, (T) sum of payouts.

Fig. 2: Experimental results for the Prisoner's Dilemma formulation.

Newcomb's Problem and Prisoner's Dilemma, RL learning algorithms learned to take actions according to the maximum expected utility solution. This is due to the fact that RL maximizes the cumulative expected reward, which is in these settings similar to the expected utility. In some situations, it might be not desirable to decide according to utility, therefore other techniques from causal decision theory are to be investigated in conjunction with learning algorithms from the field of autonomous systems.

Most existing works in the literature attempt to steer RL agents towards favourable decision equilibria by the use of modified RL algorithms. Our present results have verified that both rewarding procedure (what to reward) and the reward structure (how do we reward) are the most crucial points in shaping the agents decision. It is then beneficial to investigate further into reward shaping mechanisms or other means of rewarding. One example could be the integration of moral values, as proposed in [3]. The research in this direction could open the field of RL to a much broader philosophical discussion in decision making.

In respect of robotics, our results show how autonomous action planning in conflicting situations can be simulated. Adjusting the cooperation probabilities according to statistical data enables the simulation of different assumptions about a potential cooperation partner. The technique can be integrated into future action planning algorithms of service and maintenance robots. It is also conceivable to use it in autonomous cars to assess different traffic situations. The car would be able to simulate different outcomes of maneuvers according to statistical data for the probabilities of breaking or obeying traffic rules.

REFERENCES

[1] A. L. C. Bazzan, A. Peleteiro and J. C. Burguillo, *Learning to cooperate in the Iterated Prisoner's Dilemma by means of social attachments*, In Journal of the Brazilian Computer Society, 2011.

[2] C. Claus and C. Boutilier, *The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems*, In Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, 1998.

[3] A. Edalat, A. Ghoroghi and G. Sakellariou. *Multi-games and a double game extension of the Prisoner's Dilemma*, http://arxiv.org/abs/1205.4973, 2012.

[4] A. Flache and M. W. Macy, *Stochastic Collusion and the Power Law of Learning: A General Reinforcement Learning Model of Cooperation*, In Journal of Conflict Resolution, 2002.

[5] D. Lewis, *Prisoner's Dilemma Is a Newcomb Problem*, In Philosophical Papers (Volume II), Oxford University Press, USA, 1986.

[6] R. Nozick, *Newcomb's problem and two principles of choice*, In Essays in honor of Carl G. Hempel, Springer Netherlands, 1969.

[7] T. W. Sandholm and R. H. Crites, *Multiagent reinforcement learning in the Iterated Prisoner's Dilemma*, In Biosystems, 1996.

[8] M. Seto, *Marine robot autonomy*, Springer Science & Business Media, 2012.

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

[10] C. Thieme, I. Utne, and I. Schjølberg, *A risk management framework for unmanned underwater vehicles focusing on human and organizational factors*, 34th International Conference on Ocean, Offshore and Arctic Engineering, American Society of Mechanical Engineers, 2015.