MODELING UNLEARNING AND RELEARNING WITH MULTI-AGENT Q-LEARNING SYSTEMS

Maryam Solaiman[1], Theodore Mui[1], Qi Wang[2], Phil Mui[3]

[1]Aspiring Scholars Directed Research Program Fremont, USA
[2]University of Texas Austin Austin, Texas
[3]Salesforce, San Francisco, USA

ABSTRACT

We model unlearning by simulating a Q-agent (using the reinforcement learning Q-learning algorithm), representing a real-world learner, playing the game of Nim against different adversarial agents to learn the optimal Nim strategy. When the Q-agent plays against sub-optimal agents, its percentage of optimal moves is decreased, analogous to a person forgetting ("unlearning") what they have learned previously. To mitigate the effect of this "unlearning", we experimented with modulating the Q-learning so that minimal learning occurs with untrusted opponents. This trust-based modulation is modeled by observing opponent moves that are different from those that a Q-agent has learned. This model parallels human trust which tends to increase with those whom one agrees with. With this modulated learning, we observe that a Q-agent with a baseline optimal strategy is able to robustly retain previously learned strategy, in some cases achieving a 0.3 difference in accuracy from the unlearning model. We then ran a three-phase simulation where the Q-agent played against optimal agents in the first phase, sub-optimal agents in the second "unlearning" phase, and optimal or random agents in the third phase. We found that even after unlearning, the Q-agent was quickly able to relearn most of its knowledge about the optimal strategy for Nim.

KEYWORDS

Reinforcement learning, Q-learning, Nim Game, Unlearning, Learned Memory, Misinformation

1. INTRODUCTION

The topic of AI's alignment to human behavior is increasingly relevant today, especially with immense surges in the applications of AI, leading to questions about its ethics and conformant with human values as well as its authenticity and ability to mimic human behavior. Reinforcement learning, a machine learning model that mimics human trial-and-error processes by training a model to converge to an optimal result, is in particular being used more often, especially in the development phase of large language models, as human feedback-based reinforcement learning plays a major role in holding AI to its designers' expectations. Recently, more attention has been placed in researching how AI models can forget information, which is important in cases such as when AI needs to temporarily use and then forget private data, especially in LLMs [11]. Thus, our research, which studies how reinforcement learning (RL) models, in particular Q-learning agents, can successfully mimic human behavior when it comes

David C. Wyld et al. (Eds): NWCOM, ASOFT, ITCCMA, CSITY, BDIoT, MLNLP, SIGPRO, AIFZ– 2024
pp. 135 -143, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141912

to trust, unlearning, and learned memory, is a significant crossroad of a multifaceted approach on exploring the relationship between AI, learning, and unlearning.

Our research is centered around different agents learning to win by playing the Nim game against an opponent. The setup involves a set of piles with a given number of stones in each pile, and two players take turns removing any number of stones from any single pile. The player to remove the last stone wins the game. This game is suitable for a Q-agent system because Nim is a combinatorial game, meaning it has a known optimal strategy that can guarantee success, as well as being impartial, meaning all players have the same available moves.

The optimal strategy structure of the game holds a complementary relationship with our chosen agent for this project, a type of RL agent known as a Q-agent, aptly named because it stores its strategy in a Q-table. The strategy stored in the Q-table is the strategy it applies in the games or situations it acts in and that which it believes to be the optimal strategy for the environment. The Q-table describes the estimated total reward for taking a specific action from a specific state, across all states and actions. Based on whether it wins or loses the game, the Q-agent can update its Q-table with positive or negative rewards, with the expectation that by playing many games, the agent's Q-table can learn to converge to the optimal strategy.

Our research parallels the exigence in rising efficiency of the spread of misinformation, especially through the growing connections forged by social media, inciting a sense of urgency in our case study [1, 6, 7]. We seek to discover how this misinformation can affect individuals under different circumstances, specifically the role that bias, accuracy, and repetition play in learning information, with the hope of finding ways to mitigate this misinformation [2, 3, 4, 5, 10]. Thus, our research follows the subsequent misinformation analogy in order to make conclusions about the way we interpret misinformation and how this affects unlearning: Truth is represented by the optimal strategy to win in Nim, while individuals are represented by Q-agents with some initial beliefs trying to learn the truth by conversing, or, in this case, playing rounds of the Nim game. The games are played against one of three types of opponent agents: mal-optimal agents, optimal agents, or random agents. True news is represented by optimal agents– ie, agents which know and use the optimal strategy and do not learn or change their strategy. False or disinforming news is represented by non-learning agents such as mal-optimal agents which always use a sub-optimal strategy. Random agents' actions are generated randomly.

Our research is based upon the Bellman equation, a foundational equation in reinforcement learning, which relies upon a formula to advance the states of the Q-learning agent. Existing approaches have used the Bellman Equation on a reinforcement agent that, like ours, discerned the optimal strategy for the game of Nim [9]. Previous research has shown the importance of recognizing how machine learning models unlearn, which holds special significance with the growth of AI chatbots which may need to unlearn sensitive information [11]. The approach of using a Q-agent participating in Nim games with an analogy of misinformation was pioneered with the intent of using the results in order to form suggestions on combative strategies against misinformation in the real world [8]. However, our research advances from such procedures in that, rather than producing combative strategies against misinformation, we seek combative strategies against unlearning, which requires us to perform multi-stage simulations, in which the conditions (i.e. the adversarial agents that our learning agent trains and plays against) changes mid-simulation.

Despite the growing importance of the field, not much study has been devoted to how machine learning models, including q-learning, can mimic human memory retention. This research can provide rich dividends by advancing our understanding of how both human and machine systems

respond to situations involving misinformation, trust, and unlearning based on our experiments with the game of Nim.

Our intent with this research is to find out under which conditions reinforcement learning agents unlearn, and consequently, how to mitigate this unlearning. Furthermore, we hope to discern how much information a Q-learning agent retains after successive periods of learning and unlearning. With this information, our hope is that our research can be applied to principles of trust and learning in other machine learning systems as well as principles about the role trust plays in the spread of misinformation and how it contributes to unlearning in individuals and human populations.

2. METHODOLOGY

In each of our experiments, we used Q-learning agents playing against various predetermined, rule-based opponents in a simulation for many episodes. Our simulations all involve our Q-learning agent playing games of Nim against one of three different opponent agents: the optimal agent, which plays using the optimal strategy for the Nim game, the mal-optimal agent, which plays using a non-optimal strategy and always loses, and the random agent, which plays using a randomly generated strategy. Optimal opponents are calculated mathematically by picking stones from piles that will result in a Nim Sum of 0. Otherwise, it will randomly choose an action as all actions are optimal in a non-winning state. Mal-optimal opponents are the non-intersection between the set of all actions and the set of optimal actions.

Our Q-learning agents are all parameterized as suggested by Erik Jarleberg's thesis [9], with an exploration rate of 0.1, discount rate of 0.1, and an initial learning rate of 0.1, and the Bellman equation [Fig. 1] is used to update the Q-tables of these agents:

$$Q(s', a) \leftarrow (1 - \alpha) \cdot Q(s, a) + \alpha \cdot (r + \gamma \cdot \max_{a'} Q(s', a'))$$

Fig. 1 Bellman Equation for Q-learning agents

By adding a modifier that subtracts the predicted Q-value of the future state, we improved the convergence rate for the Q-learner.

Ideally, after a number of simulated games, the Q-learning agent will learn the strategy for Nim and converge to the optimal strategy.

We ran two sets of simulations: The first set involved two-phase simulations, in which we had a learning phase followed by an unlearning phase. The second type involved a three-phase simulation containing a learning phase followed by an unlearning phase and then a relearning phase.

Our first experiment involved two-phase simulation in which the first phase, consisting of 70,000 simulations, involved simulated games played only against the optimal agents, and the second half, another 70,000 simulations, were played only against mal-optimal agents. Due to this, the accuracy drops from 1.0 in the first half to 0.6 in the second half. This significant drop in accuracy occurs because the Q-agent rewrites its Q-table– a phenomenon known as unlearning.

To mitigate this, we used a Modified Learning Rate strategy, in which the Q-agent judges its opponent's actions against its own Q-table. If the opponent's action is not the action the Q-agent

would perform in that state, the learning rate is scaled by a multiplier of 0.1, thereby modeling a greater trust given to similar agents.

Our second experiment was modeled similarly to the first, except the first learning phase involved the optimal agent as the primary opponent agent, and the second phase involved random agents.

The third experiment is distinct from the first two in that the learning phase involves a suboptimal strategy–the random opponent agent– in the learning phase and the optimal strategy in the unlearning phase to test whether unlearning occurs, ie, whether the convergence rate was affected by the first phase.

The fourth experiment involved a three phase simulation. The first phase (learning phase) consists of 5,000 rounds where the Q-agent plays against optimal or random agents. The second phase (unlearning phase) consists of 10,000 rounds where the Q-agent plays against mal-optimal agents. The third phase (relearning phase) consists of 5,000 rounds where the Q-agent plays against optimal or random agents again, in order to discern how much of the strategy is retained from the first phase.

3. RESULTS

3.1. Experiment 1: Learning with Optimal Agents, Unlearning with Mal-optimal Agents

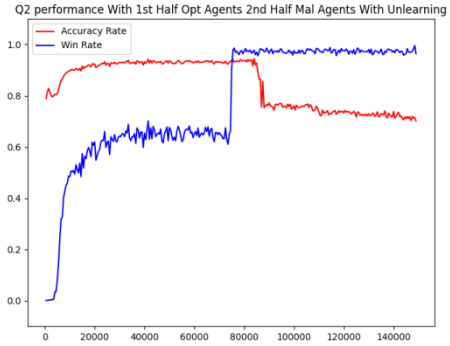

Fig. 2 Unlearning with Mal-optimal Agents without the Modified Learning Rate Strategy

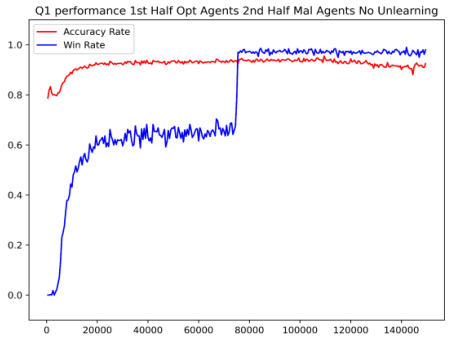

Fig. 3 Unlearning with Mal-optimal Agents with the Modified Learning Rate Strategy

Without the Modified Learning Rate strategy, the mal-optimal agents cause the Q-agent's accuracy to drop from 1.0 (its accuracy during phase 1, when it played only optimal agents) to 0.7 (its accuracy during phase 2 when it plays only mal-optimal agents) as it rewrites its Q-table information, signifying unlearning [Fig. 2].

Using the Modified Learning Rate Strategy, by giving less trust to differing opponents, thus judging them, the Q-agent achieves a stable accuracy rate of 1.0 throughout the entire simulation, even during phase 2, when playing with solely mal-optimal agents [Fig. 3].

The win rate remained identical for both simulations because optimal agents usually win the game of Nim because they are built on the optimal strategy, and the mal optimal agents always lose.

3.2. Experiment 2: Learning with Optimal Agents, Unlearning with Random Agents

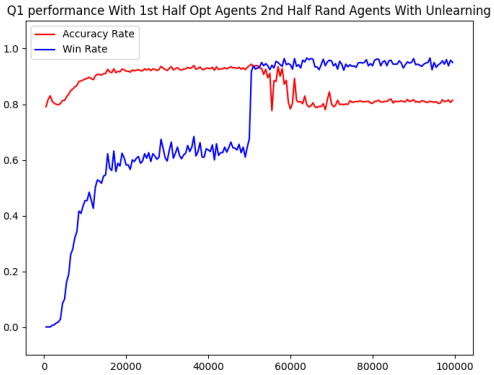

Fig. 4 Unlearning with Random Agents without the Modified Learning Rate Strategy

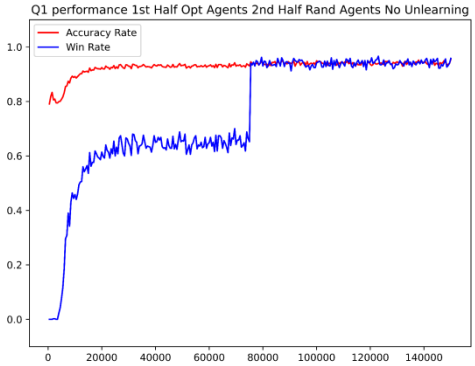

Fig. 5 Unlearning with Random Agents with the Modified Learning Rate Strategy

For the first half of its training period, the agent played against an optimal agent, resulting in a high accuracy of around 1.0 and a low win rate. In the first simulation [Fig. 4], without the Modified Learning Rate Strategy, accuracy dropped to around 0.8 during the unlearning phase. With the Modified Learning Rate strategy, the Q-agent was able to sustain its accuracy of 1.0 throughout the unlearning period [Fig 5]. Since it was playing against a random agent, it also had a relatively high win rate.

The main difference between playing against the mal-optimal agent (Experiment 1) and playing against the random agent (Experiment 2) was that when playing against the random agent, the win rate was exactly the same as the accuracy rate.

This is due to epsilon, or the exploration rate, which introduces a point of randomity that causes the win rate during random agent games to be only as high as the agent's moves themselves. This is not a problem when playing against the mal-optimal agent, as it always loses.

3.3. Experiment 3: Learning with Random Agents, Unlearning with Optimal Agents

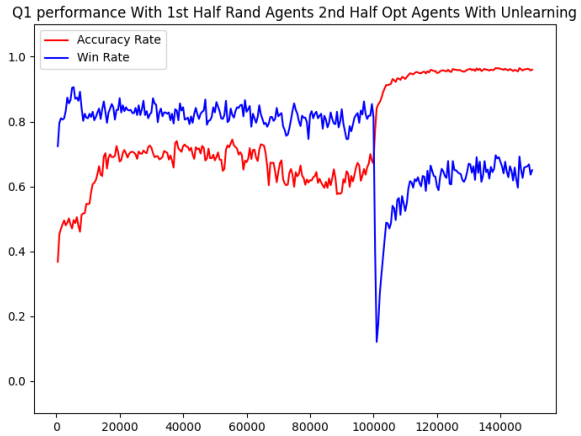

Fig. 6 Learning with Random Agents, Unlearning with Optimal Agents without the Modified Learning Rate Strategy

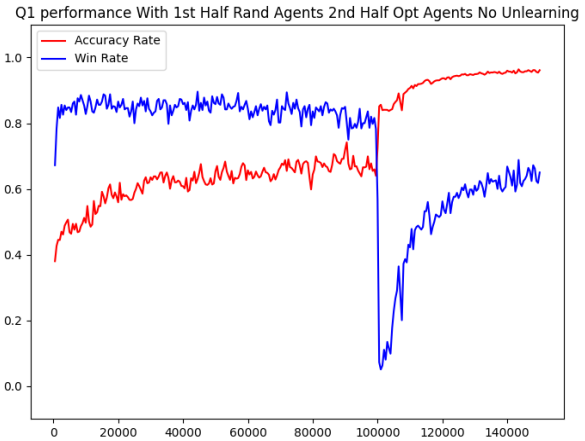

Fig. 7 Learning with Random Agents, Unlearning with Random Agents with the Modified Learning Rate Strategy

For the first half of its training period, the agent played against a random agent, resulting in a low accuracy of around 0.6 and a win rate of around 0.8 [Fig. 6].

For the second half of this training period, the agent played against a random agent but scaled the learning rate by a multiplier of 0.1 if the agent's move for a given state did not match with the Q-agent's moves [Fig. 7]. Despite this uneven scaling for agents that have different suggested moves, the agent was able to learn from the optimal agent because the opponent agent always won, forcing the Q-agent to update its Q-table.

Since it was playing against an optimal agent, the Q-agent's win rate was around 0.6 after originally plummeting, but its accuracy climbed to around 1.0. It learned slower using this strategy, but it still learned the optimal strategy.

3.4. Experiment 4: Unlearning and Relearning in Three Phases

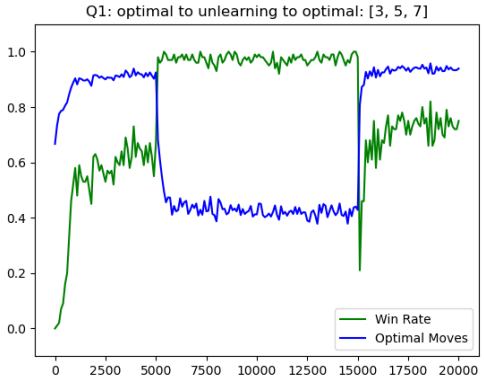

Fig. 8 Phase 1 Optimal, Phase 2 Unlearning, and Phase 3 Optimal

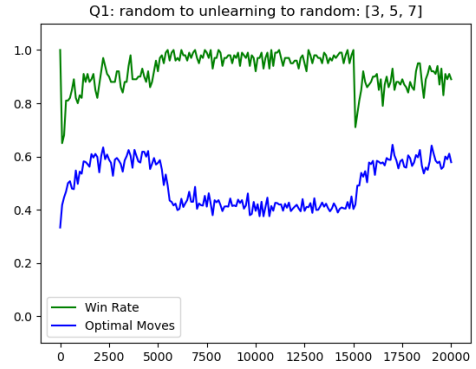

Fig. 9 Phase 1 Random, Phase 2 Unlearning, and Phase 3 Random

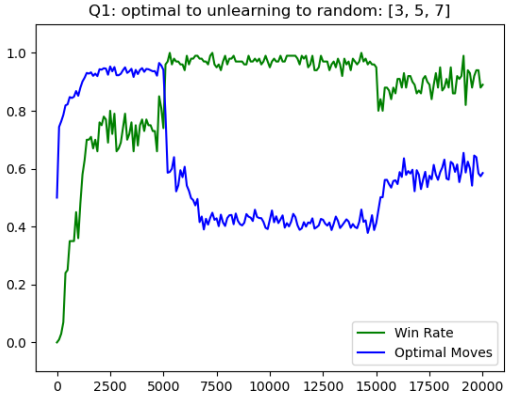

Fig. 10 Phase 1 Optimal, Phase 2 Unlearning, and Phase 3 Random

Three scenarios appear in the figures above: optimal to unlearning to optimal [Fig. 8], random to unlearning to random [Fig. 9], and optimal to unlearning to random [Fig. 10]. In the optimal to

unlearning to optimal simulation and the random to unlearning to random simulation, the Q-agent is able to return back to the exact same accuracy in the third phase as the first phase. In the optimal to unlearning to random simulation, the Q-agent goes from an accuracy of 0.95 in the first phase to an accuracy of 0.4 in the second phase to an accuracy of 0.6 in the third random phase, signifying a retention of some accuracy, even though it is not the full original accuracy. These results show that the Q-agent retains its knowledge of the optimal strategy even after unlearning when it is re-exposed to a didactic agent.

4. CONCLUSION

Our model represents misinformation with a novel multi-agent Nim game and using Q-learning for modeling learning behaviors. Current findings provide proof of concept that Reinforcement Learning agents do accurately simulate human actions.

In our experiments, the Modified Learning Rate strategy proved successful in mitigating unlearning. Thus, we conclude that when simulated learners are more likely to learn from those that are more similar to them, they can better retain their knowledge of an optimal strategy even in the presence of disinformation agents. This sense of "trust" shields them from the misinformation caused by non-optimal agents and shows a mitigation of unlearning, which can be extended to emphasize the role of trust in mitigating the effect of misinformation in human populations. Without the sense of trust provided in the Modified Learning Rate strategy, however, the agent was able to unlearn, providing grounds to reverse this principle of trust when desired in other machine learning systems to initiate unlearning.

Furthermore, our Q-agent learned the optimal strategy even from games against random or mal-optimal agents. Thus, Q-learning agents can often learn from those agents that are different from them even if they judge against them, albeit at a slower rate. When paralleled in human populations, we see that slowed learning occurs in the presence of those with differing opinions.

When simulated Q-learners are reintroduced to either random or optimal agents after a hiatus of about 10,000 episodes, they are able to pick up and relearn the optimal Nim strategy. This indicates learned memory, showing that despite unlearning, the learned optimal strategy is stored and revealed when re-exposed to optimal agents. Furthermore, the span of the unlearning period exceeded that of the learning period, revealing that increased exposure to unlearning agents does not affect the aforementioned learned memory.

Thus, our findings reveal the role of trust in mitigating unlearning or misinformation, show that even non-optimal agents can result in the learning of the optimal strategy in Q-agents, and exemplify learned memory. When these findings are applied to human populations, they could provide strategies to combat misinformation and unlearning on the individual level by introducing the role of trust in learning, validating the effect of disinformation on learning the truth, and showing that unlearning can be reversed by a reintroduction of information. When applied to other machine learning systems, our findings could provide strategies on propelling unlearning if necessary by introducing disinformation to undo the learning or forget the information.

REFERENCES

[1] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, ”The spreading of misinformation online,” Proceeding of the National Academy of Science (PNAS) vol. 113 (3), pp. 554-559, January 2016.

[2] J. L.Foster, T. Huthwaitea, J A.Yesberg, M. Garry, E. F. Loftus, ”Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses, Acta Psychologica, vol. 139, pp. 320-326, February 2012.

[3] J. B. Bak-Coleman, I. Kennedy, M. Wack, A. Beers, J. S. Schafer, E. S. Spiro, K. Starbird J. D. West, Combining interventions to reduce the spread of viral misinformation, Nature Human Behaviour, 2022

[4] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Bernisky, et al., ”The science of fake news,” Science, vol 359, No. 6380, pp. 1094–1096, March 9, 2018.

[5] G. Pennycook, Z. Epstein, M. Mosleh, A. Arechar, D. Eckles, D. G. Rand, ”Shifting attention to accuracy can reduce misinformation online.” Nature 592, 590–595, 2021.

[6] D. Klepper, “’Horrifying’ Conspiracy Theories Swirl around Texas Shooting.” AP News, Associated Press, 26 May 2022.

[7] B. Lewis. “All of YouTube, Not Just the Algorithm, Is a Far-Right Propaganda Machine.” Medium, FFWD, 9 Jan. 2020.

[8] E. Huang, P. Mui. ”Modeling Misinformation With Q-Learning, Nim, and Multi-Agents,” 2022 IEEE MIT Undergraduate Research Technology Conference (URTC), 2022.

[9] E. Järleberg, “Reinforcement Learning on the Combinatorial Game of Nim.” KTH Computer Science and Communication, 04-2011.

[10] C. G. Luca. “Biases Make People Vulnerable to Misinformation Spread by Social Media.” Scientific American, 21 June 2018.

[11] H. Xu, T. Zhu, L. Zhang, et al., “Machine Unlearning: A Survey.” ACM Computing Surveys, June 2023.