**Markov Decision Processes (MDPs)**

Soda bottles are major environmental pollutants affecting nearly every country in the world. Studies have shown more than 3 million soda bottles are recklessly disposed to the environment every day without proper waste management. Although Coca Cola through its global agencies has tried to regulate and recycle the use of soda bottles, waste management of soda bottles is still a major problem riddling the company. Many environmental organizations have decried about the danger posed by soda bottles. According to many global environmental agencies, such as Intergovernmental Panel on Climate Change (IPCC), United Nations Environment Program (UNEP), and European Environment Agency (EEA), soda bottles pose serious threats to fauna and flora. To end this menace, Recycling Robot (RR) is recommended. RR is recommended because it uses artificial intelligence to collect and dispose of soda bottles in recycling bins.

**Recycling Robot’s Markov Decision Processes (MDPs)**

In this paper, Recycling Robot (RR) is used to illustrate the application of MDPs to solve a serious problem affecting the environment. RR uses sensors and artificial intelligence to locate soda bottles that are not disposed of properly, and then put them in a recycle bin. Artificial intelligence (AI) is used by RR because it offers accurate and real-time information that helps the robot to identify the location of soda bottles, do sorting, and put them in the right recycling bins (Kumar & Varakantham, 2017). However, different states (S) compel the RR to take actions (A) to ensure its continuity and to manage the power to avoid battery depletion or chances of failed operations. Also, the actions of RR are determined by external factors. These external events are dictated by the power status and rewards ®. Moreover, the probability of RR moving to the new state (S\*) is depended on the corresponding rewards (Ra) as well as the transition function (Pa).

**States of RR**

In this case, the RR is in two states; low and high charge levels. The state set is S, (High, Low). The Low represents the Low battery level or a state where the battery is about to be depleted, or where RR has limited time to power the RR. On the other hand, High represents the low battery level or a state where will take more time to get depleted.

**Actions in Each State**

In each state, RR can take three actions (A), (1) search for cans for a fixed amount of time (searching the cans actively), (2) remain stationary and wait for someone to bring the cans, or (3) go to charging station (home base) to its battery. However, recharging is only allowed when the battery is at a low state because recharging is pointless or meaningless when the battery level is high (Suárez, Torras & Alenyà, 2019). In these three actions (options), searching the cans is the best way to identify, locate, and find cans. Notwithstanding, searching the cans actively can deplete the RR’s battery. On the other hand, waiting for someone to pick the cans cannot run down the battery.

Whenever searching the cans actively, the highest possibility is that the battery level will become drained or depleted. When the battery is depleted, the RR shuts down and waits to be charged (triggering a negative reward). Under BB’s MDPs, it is assumed that the decisions are entirely the function of the battery level. These two levels are low and high, and the state set is S (High, Low).

The action (A) is (wait, search, recharge). Based on the state-dependent action, the two sets are; A(high)= (search, wait), and A(low)= (wait, search, recharge). In A(low), the recharge is possible (included) because the battery is depleted. In A(high), the recharge is meaningless (excluded) because the battery is full. Searching for cans when the energy level is high might not change the state the probability a. However, when the energy level is low, the expected probability is1-a. On the other hand, searching for cans when the energy level is low at probability b and drains the battery with probability (1-b). At the A (low), RR must be rescued by recharging back to a high energy level.

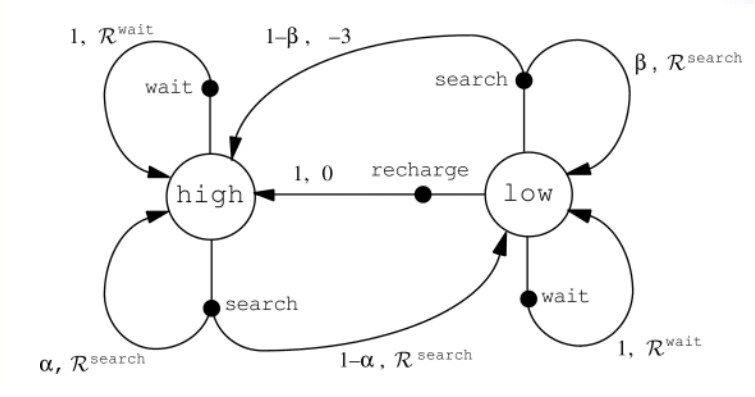
Assuming that each can be collected by RR represents a unit reward, +10 depicts that RR found 10 cans (10 rewards). However, if the RR chooses to wait, it will not collect any can, thus the reward is 0. On the other hand, rewards of -3 are realized when the robot has been recharged or rescued. Also, R-search represents the number of cans collected by the RR when in research mode. R-wait represents the number of cans collected by RR when in wait mode.

**Transition Function**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S=St | S\*=st+1 | a=a(t) | Pa (S, S\*) | Ra (S, S\*) |
| High | High | Search | a | R-search |
| high | Low | Search | 1-a | R-search |
| Low | High | Search | 1-b | -3 |
| Low | Low | Search | B | R-search |
| High | High | Wait | 1 | R-wait |
| High | Low | Wait | 0 | R-wait |
| Low | High | Wait | 0 | R-wait |
| Low | Low | Wait | 1 | R-wait |
| Low | High | Recharge | 1 | 0 |
| Low | Low | Recharge | 0 | 0 |

**Fig 1**

In figure 1, the S represents the state, S\* represents the new state, Ra (S, S\*) is the corresponding rewards, and Pa (S, S\*) is the transition function. The probability that this process moves into S\* is dictated by the action taken, which is represented by the transition function Pa (S, S\*). The next state S\* is entirely depended on the S, and the action of decision-maker a. However, given that the a and s are independent, they become conditionally independent of past actions and states (Oner et al., 2019). In other words, the transition through transition function satisfies the Markov property. Although it is often perceived that MDPs are synonymous with Markov chains, they differ because of allowing choices (additional actions) and giving motivation (rewards). Fig 2 summarizes the information above.



**Fig 2**

**Reward Function (Ra)**

In fig 1, the expected reward function Ra (S, S\*) is shown for all the states (Low, High) and actions (search, wait, recharge). For example, the Ra for (Low, High, Search) and a Pa (1-b) is -3. This implies that the rewards are negative, or rewards realized when the robot has been recharged or rescued. The Ra for (Low, High, Recharge) and a Pa (1) is 0. Similarly, the Ra for (Low, low, Recharge) and a Pa (0) is 0.

**Optimal Value Function and Recommended Policy**

Based on the action-value function for policy Ω arising from the value of taking action a, the value of the action is Q Ω (s, a), the optimal value is St=1\_Pa (St, St, =1), where αt= Ω. Therefore, the optimal value is 1+0=1. The optimal value of 1 depicts that the policy (Ω) is moderately effective in addressing the issue caused by poor plastic management. In other words, the value of 1 suggests that the policy possesses a uniform convergence rate for horizontal rewards under the assumptions presented above (Chen, Guo & Bai, 2017). However, other hidden variables including inflation, taxation, and adoption level of the policy, are likely to influence either directly or indirectly the policy.

In this case, the recommended policy is to optimize the search for cans using the RR. Although searching for cans might deplete the battery, it is the best action because it guarantees optimal value. To enhance the effectiveness of the search option, it prudent to improve the research systems used. Alternatively, a backup battery should be used. Using backup batteries is prudent because it minimizes the time wastage in recharging RR’s battery. Also, it eliminates the cost associated with the recharging process.

Considering that issue caused by soda bottles is immense, there is a need for policymakers to enact laws that compel Coca-Cola and its subsidiaries to use eco-friendly bottles or use RR under that actively searching option. For many years, Coca-Cola and other beverage companies have been accused of failing to utilize RR because they only focus on profit maximation at the expense of the environment. Although Coca-Cola claims that it has undertaken great measures through corporate social responsibility (CSR), it has failed to embrace the latest technology for waste management.

**Conclusion**

Utilizing Markov Decision Processes (MDPs) for RR is very critical in identifying actions, states, and transition and reward functions for effectively managing the issue posed by soda bottles. Out of the three actions, it is clear that searching the cans actively is the best option. It is the best because it assures continuous management of soda bottles. Therefore, policymakers should create laws and policies that compel these companies to leverage on RR or similar alternative solutions for waste management.

**References**

Chen, Q., Guo, X., & Bai, H. (2017). Semantic-based topic detection using Markov decision processes. *Neurocomputing*, *242*, 40-50.

Kumar, R. R., & Varakantham, P. (2017, May). Exploiting Anonymity and Homogeneity in Factored Dec-MDPs through Precomputed Binomial Distributions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (pp. 732-740).

Oner, A., Albayrak, M. S., Guner, F., & Atakli, I. M. (2019, June). An Activity Recognition Application Based on Markov Decision Process Through Fish Eye Camera. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 251-258). IEEE.

Suárez-Hernández, A., Torras, C., & Alenyà, G. (2019). Practical Resolution Methods for MDPs in Robotics Exemplified with Disassembly Planning. *IEEE Robotics and Automation Letters*, *4*(3), 2282-2288.