Somatic Q-Learning

Artur P. Carneiro[1], **Danilo H. Perico**[1], **Reinaldo A. C. Bianchi**[1]

[1]Centro Universitário da Fundação Educacional Inaciana - FEI
São Bernardo do Campo, SP, Brasil

acarneiro@fei.edu.br, dperico@fei.edu.br, rbianchi@fei.edu.br

Abstract. *Reinforcement Learning (RL) is an area of machine learning that utilizes algorithms inspired by biological concepts, where an agent learns from the actions it takes, the resulting states, and the rewards obtained from the environment. In this area, one of the most used algorithms for off-model environments is Q-Learning. This algorithm has some limitations regarding the number of possible actions and the size of the state space, in addition to an exponential increase in training time related to these two variables, making some applications of it unfeasible. This research presents an adaptation of the algorithm, utilizing a mechanism inspired by the functioning of somatic markers, as proposed by António Damásio, to enable the use of Q-Learning in environments where it is infeasible.*

Resumo. *O Aprendizado por Reforço é uma área do aprendizado de máquina com algoritmos inspirados em conceitos biológicos, onde um agente aprende a partir de uma ação e a observação do estado e da recompensa, resultantes desta ação, obtidos do ambiente. Nesta área, um dos algoritmos mais utilizados para ambientes não modelados é o Q-Learning. Este algoritmo possui algumas limitações quando ao número de ações possíveis e tamanho do espaço de estados, além de um aumento exponencial do tempo de treinamento em relação à essas duas variáveis, tornando algumas aplicações deste inviáveis. Esta pesquisa apresenta uma adaptação deste algoritmo, utilizando um mecanismo bio-inspirado no funcionamento dos marcadores somáticos, proposto por António Damásio, com o objetivo de viabilizar o uso do Q-Learning para ambientes onde este mostra-se inviável.*

1. Introdução

Inspirado no funcionamento do aprendizado biológico, os algoritmos de RL revolucionaram a capacidade de sistemas computacionais tomarem decisões autônomas em ambientes complexos, desde jogos eletrônicos até a robótica. Tradicionalmente, esses algoritmos baseiam-se em análises racionais de utilidade esperada, seguindo princípios cartesianos que priorizam maximizar recompensas futuras. No entanto, em cenários altamente dinâmicos, com alta dimensionalidade de estados e ações, o RL enfrenta desafios críticos: alto custo computacional, lentidão no treinamento e eficiência limitada em situações inéditas. Essas mesmas dificuldades não são observadas em seres vivos que se utilizam da mesma análise racional que serviu de inspiração para o RL.

Na neurociência, o trabalho do neurologista António Damásio apresenta a Hipótese do Marcador Somático (HMS), propondo que emoções e sensações

corporais, registradas como “marcadores” ao longo da vida, atuam como atalhos cognitivos, eliminando opções inviáveis e acelerando decisões. Em humanos, esse mecanismo permite equilibrar racionalidade e intuição, otimizando escolhas sob incerteza.[Damásio 1994] Desta forma, além da análise racional, a tomada de decisão biológica conta com os marcadores somáticos para otimizar seus processos.

Neste contexto, este artigo propõe um modelo de Aprendizado por Reforço Somático, adicionando ao RL um mecanismo inspirado no funcionamento dos marcadores somáticos. Essa abordagem busca:

1. Reduzir a exploração desnecessária de estados/ações, filtrando opções com base em experiências emocionais prévias.
2. Acelerar a convergência do treinamento, priorizando ações marcadas como positivas.
3. Melhorar a adaptabilidade em ambientes dinâmicos, mesmo com estados desconhecidos.

Para demonstrar o funcionamento deste modelo, desenvolve-se uma adaptação do algoritmo de Q-Learning, chamado de Q-Learning Somático e realiza-se uma comparação entre dois agentes: um Q-Learning e outro Q-Learning Somático, a fim de avaliar as características citadas.

Este artigo detalha a arquitetura do modelo, algoritmos e os resultados comparativos, abrindo caminho para agentes mais eficientes e biologicamente inspirados.

2. Conceitos Fundamentais

2.1. Aprendizado por Reforço

O RL é inspirado no processo de raciocínio biológico onde os seres dotados deste são capazes de, com base nas experiências desenvolvidas ao longo da vida, tomar decisões sobre sua existência. No RL, o aprendizado ocorre por recompensas. O agente interage com o mundo e, a cada interação, recebe recompensas, chamadas na terminologia da psicologia de “reforços”. Essas recompensas indicam ao agente como ele está indo ou o quão perto está do objetivo [Russell and Norvig 2021]. A recompensa é sempre obtida pela observação do ambiente no estado atingido após a execução de uma ação.

Nesse sentido, o conceito de aprendizado por reforço pode ser formalizado utilizando a teoria de Processo de Decisão Markoviano (MDP - *Markov Decision Process*). O MDP é “um problema sequencial de decisão para um ambiente totalmente observável, estocástico com um modelo de transição Markoviano e a adição de recompensas” [Russell and Norvig 2021]. O modelo de transição Markoviano é definido por $P(s'|s,a)$ - a probabilidade de se atingir um estado s' dado um estado inicial s e uma ação a. A recompensa recebida é definida por $R(s,a,s')$ - valor da recompensa por, estando no estado s, executar a ação a e atingir o estado s' - podendo ser positivo ou negativo e estando sempre no intervalo de $\pm R_{max}$. Desta forma, o objetivo de um agente markoviano é maximizar a somatória das recompensas recebidas. O objetivo de um agente de RL é o mesmo, porém, enquanto o agente markoviano tem o MDP como problema à resolver o agente RL “está em” um MDP. Ele pode não saber qual o modelo de transição ou a função de recompensa, mas ele tem que realizar a ação para aprender.

2.2. *Q-Learning*

O Q-learning foi apresentado por Chris Watkings em 1989, em sua tese de doutorado *Learning from Delayed Rewards* pela universidade de Cambridge. O propósito é "aprender" uma "função de utilidade de ação" $Q(s, a)$, onde s é um estado do universo de estados e a uma ação do universo de ações. Essa função substitui a utilidade esperada $U(s)$ e é utilizada para se chegar à melhor política π^*, dado um estado s, uma vez que:

$$U(s) = \max_a Q(s, a) \tag{1}$$

e

$$\pi^*(s) = \arg\max_a Q(s, a) \tag{2}$$

Desta forma "$Q(s, a)$ denota a recompensa total esperada se o agente tomar a ação a em s e agir de forma otimizada depois disso"[Russell and Norvig 2021]. Um agente Q-learning então navega pelos estados experimentando as ações disponíveis, percebendo as recompensas e desenvolvendo uma "tabela" (s, a). De maneira a considerar os valores dos estados adjacentes e o aspecto temporal das experiências do agente, utiliza-se a chamada equação de Bellman para atualização temporal de $Q(s, a)$ em cada iteração do agente, compondo o que chamamos de TD (*Temporal-difference*) Q-learning:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[R(s, a, s') + \gamma \max_{a'} Q(s', a') - Q(s, a)] \tag{3}$$

onde:

- α - taxa de aprendizado (intervalo $(0, 1]$)
- γ - fator de desconto do erro (intervalo $(0, 1]$)
- R - recompensa

2.3. Hipótese do Marcador Somático

Em 1994, Damásio descreve a hipótese da existência de um mecanismo chamado por ele de "marcador somático", responsável por marcar (positivamente ou negativamente) experiências somáticas (sensações, emoções e até pensamentos). Essas marcações são utilizadas pelos seres delas dotadas como parâmetro nas tomadas de decisões e planejamento. Nesta hipótese, apresenta um processo decisão em que, antes de qualquer análise das opções ou antes de qualquer raciocínio, o marcador somático, a partir do estado identificado, recupera (caso exista) uma memória "somática" (do grego, soma significa corpo) e faz o indivíduo "sentir" as sensações que emergiram no momento que aquele estado foi "marcado". Essa função de recuperação de sensações, chamada de "como-se", faz convergir a atenção do indivíduo para resultado (positivo ou negativo) das ações já tomadas no passado quando desse mesmo estado (experiência). Esse sinal pode fazer com que o indivíduo rejeite imediatamente essa ação, eliminando-a das possibilidades, ou fazer com que aceite imediatamente uma ação positiva. Caso não sejam de intensidades altas o suficiente (para rejeitar ou aceitar uma ação somaticamente), ainda assim, essas sensações interferem na análise racional, reduzindo ou amplificando o quão vantajosa pode ser uma ação. Ou seja, o marcador somático pode não eliminar a execução do processo de raciocínio subsequente, porém, ele pode reduzir drasticamente o número de opções a serem analisadas ou ainda aumentar a precisão do processo decisório.

3. Trabalhos Relacionados

Desde a proposição da HMS, pesquisadores desenvolveram mecanismos na tentativa de sintetizar a estrutura proposta por Damásio em agentes computacionais para tomada de decisão. Um primeiro destaque vai para o trabalho de [Maçãs et al. 2001] que apresentam o DARE (Desenvolvimento de Agente Robótico baseado em Emoções). Nesse trabalho eles apresentam uma arquitetura para um agente utilizando as diretrizes apresentadas por Damásio para a análise dos estímulos do ambiente por um corpo.

Mais tarde, [Pimentel and Cravo 2009] propõem um agente com um mecanismo de marcador somático capaz de decidir qual ação tomar a partir das preferências somáticas e racionais dessa ação. [Hoefinghoff and Pauli 2013] propõem um mecanismo de aprendizado reverso baseado exclusivamente em marcadores somáticos. Nesse caso, não existe componente racional na tomada de decisão do agente. [Cominelli et al. 2018] apresentam uma adaptação de uma plataforma para robôs sociais já existente chamada SEAI (*Social Emotional Artificial Intelligence* - Inteligência Artificial Social Emocional). SEAI é um sistema cognitivo para robôs sociais e emocionais bio inspirado na cognição humana, altamente modular e com capacidade de raciocínio de alto nível. Eles desenvolveram modificações e incrementos nos módulos desse sistema para adicionar o mecanismo de marcador somático proposto por Damásio ao seu processo cognitivo.

[Cabrera-Paniagua et al. 2023] desenvolvem um sistema autônomo para tomada de decisões sobre qual trajeto um transporte de passageiros deve tomar, dado uma origem e um destino. O sistema proposto considera uma variável psicossomática que reflete a experiência do passageiro em operadores de transporte do mesmo tipo, considerando variáveis racionais e a experiência "somática" particular do passageiro.

3.1. Diferenciais e Contribuições

Os trabalhos anteriores possuem características em comum com a proposta aqui desenvolvida, como o Marcador Somático atuando como fator na tomada de decisão ([Cominelli et al. 2018] e [Cabrera-Paniagua et al. 2023]), o Marcador Somático como gatilho para decisão [Pimentel and Cravo 2009], fatores temporais na influência do Marcador Somático ([Cominelli et al. 2018] e [Cabrera-Paniagua et al. 2023]) e consideração de estados internos do agente nos marcadores ([Maçãs et al. 2001] e [Cominelli et al. 2018]). Porém todos os trabalhos, individualmente, atuam sobre partes dessas características apenas ou, quando são mais abrangentes, as implementações são específicas, com objetivos e plataformas de atuação restritos.

Este trabalho se distingui destes, buscando:

- Propor um mecanismo genérico e escalável, aplicável a diferentes algoritmos de RL.
- Considerar os marcadores somáticos nas suas duas atribuições no processo decisório: como gatilho de decisão imediata e como fator de influência na ponderação racional.
- Validar experimentalmente em ambientes simulados, demonstrando, por meio de comparação entre agentes computacionais, as características almejadas.

4. Aprendizado por Reforço Somático

Inspirado na HMS, apresentamos um modelo que almeja complementar qualquer algoritmo de RL com o mecanismo de marcador somático, chamado de Aprendizado

por Reforço Somático (*Somatic Reinforcement Learning* - SRL), como visto na Fig. 1. A estrutura relativa ao mecanismo de marcador somático é destacada nos objetos em verde. Consiste em 2 novos componentes: *Sentir* e *Avaliação Somática*, além de um conjunto de emoções artificias e uma memória somática. O componente Sentir captura o estado e a recompensa do Ambiente. Este componente executa quatro etapas internas:

1. registra o marcador somático relativo às emoções disparadas como consequência da ação tomada no estado anterior na memória somática
2. recupera da memória somática as ações (e suas respectivas emoções) marcadas anteriormente relativas ao estado atual
3. avalia se a ação com sentimento mais positivo das ações recuperadas supera um limite de agradabilidade. Se o sentimento superar, a ação é executada instantaneamente (marcação 1 da Fig. 1). Neste caso, somente o mecanismo de marcador somático atua, excluindo a execução do algoritmo de aprendizado por reforço neste ciclo.
4. caso nenhuma ação seja executada na etapa anterior, gera um conjunto de Respostas Somáticas para o estado atual.

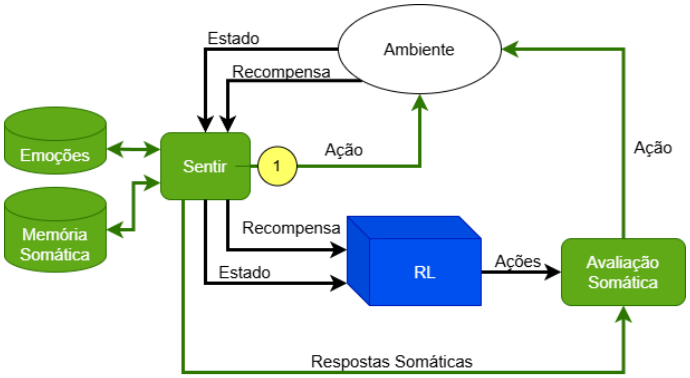

Figura 1. Aprendizado por Reforço Somático (fonte: Do Autor)

O repositório de *Emoções* é formado por funções computacionais responsáveis por sintetizar emoções no agente referentes ao estado atual, anterior, ação tomada e recompensa, podendo também considerar estados internos do agente no seu processamento. Essas funções retornam um valor numérico real entre -1 e +1, que representa o *sentimento* desta emoção. O valor zero representa uma indiferença com relação à esse sentimento. Na primeira etapa do componente *Sentir*, todas as emoções são disparadas em cada interação com o ambiente. A sensação resultante de uma ação executada em um estado é, então, composta pela média de todas as sensações das emoções disparadas, ignorando-se as sensações com valor zero (indiferentes).

O repositório de *Memória Somática* armazena os marcadores somáticos compostos pelo valor do sentimento resultante para cada par estado/ação. A criação/atualização destes valores é realizado também na etapa 1 do componente *Sentir*, após o disparo das emoções. No caso de criação de uma nova memória somática, o valor do sentimento resultante integral é armazenado. No caso de atualização de uma memória somática já existente, existe um fator de decaimento, responsável por ponderar a relevância de um novo valor perante o valor já existente na memória. Assim temos:

$$MarcadorSomatico_{e,a} \Leftarrow (1-d) \times MarcadorSomatico_{e,a} + d \times S \quad (4)$$

Onde:

- $MarcadorSomatico$: memória dos marcadores somáticos
- d : fator de decaimento da memória somática
- e : estado
- a : ação
- S : valor resultante do sentimento

Na segunda etapa do componente *Sentir*, as ações e seus respectivos sentimentos para o estado atual do agente, são recuperadas da *Memória Somática*. Essas ações e valores são avaliadas nas etapas seguinte, onde, num primeiro momento, destaca-se a ação com o maior valor de sentimento. Esta é comparada com um parâmetro de limite de agradabilidade (la). Esse parâmetro simboliza o valor limite de sentimento que deve ser interpretado como agradável. Caso a ação destacada ultrapasse esse valor, ela será executada pelo agente. No momento seguinte, caso nenhuma ação seja executada, monta-se um conjunto de informações que vão alimentar o componente *Avaliação Somática*, chamado de *Respostas Somáticas*. Essas respostas são compostas por dois sub-conjuntos: as ações proibidas (APs) e as ações marcadas(AMs). As APs são as ações constantes no conjunto das recuperadas da memória somáticas cujo valor de sentimento é inferior ao parâmetro de limite de desagradabilidade (ld). Esse parâmetro indica um valor de sentimento que deve ser considerado como desagradável. As AMs são todas as demais ações (não proibidas) com seus respectivos valores e sentimento.

Caso o componente *Sentir* não tenha executado nenhuma ação, as informações de estado e recompensa são alimentadas no mecanismo de RL subsequente, que será executado normalmente, porém, em vez de executar a ação de maior valor avaliada, o conjunto de ações e seus respectivos valores avaliados é alimentado no componente de *Avaliação Somática*. Esse componente, então, remove as ações do APs do conjunto de ações recebido e executa uma avaliação da Utilidade Esperada (UE) para cada ação do conjunto de ações possíveis restantes do RL, considerando também as Respostas Somáticas, utilizando a seguinte equação:

$$UE(a) = \frac{FV(a) + (AMs(a) * f)}{2} \quad (5)$$

Onde:

- $UE(a)$: valor da utilidade esperada para a ação a
- $FV(a)$: Função Valor para uma ação a
- $AMs(a)$: valor do conjunto de Ações Marcadas das Respostas Somáticas para uma ação a
- f : parâmetro de normalização das sensações

O parâmetro de normalização f é utilizado para equalizar possíveis diferenças entre a escala do valor de sensação e a escala dos valores da função valor de cada ação.

Calculados os valores de utilidade esperada para cada uma das ações possíveis, o *Avaliação Somática* executa a ação de maior valor avaliado.

Caso, após a remoção das ações do conjuntos APs não reste ações possíveis, o agente não vai executar a escolha somática da ação e selecionará a ação conforme o RL propõe, ou seja, seleciona a ação de maior valor do conjunto de ações recebidas.

5. Q-Learning Somático

Para este estudo, esse modelo é aplicado ao algoritmo de RL *Q-Learning*. O algoritmo foi construído como adaptação do Q-Learning citado por [Sutton and Barto 2018], e atua como componente RL do modelo.

A função $EscolherAcao$ é atribuição do componente *Avaliação Somática* e a $Sentir$, do componente *Sentir* do modelo. MS representa o repositório de memória somática.

Das linhas 7 a 15 do algoritmo do Q-Learning Somático (Algoritmo 1) temos o aprendizado somático, onde os marcadores são registrados após o disparos dos sentimentos. Função essa também do componente *Sentir* do modelo.

Algoritmo 1: Algoritmo para o treinamento do *Q-Learning* Somático

```
   Entrada:
   α ∈ (0, 1]
   γ ∈ (0, 1]
1  Inicializar Q(s, a) para todo s ∈ S+, a ∈ A(s), arbitrariamente com exceção de Q(terminal, .) = 0
2  para cada episódio faça
3      Inicialize s
4      enquanto s não terminal faça
5          a ← EscolherAcao
6          Execute a ação a e observe r, s'
7          F ← Sentir(s, r, s', a)
8          se F ≠ 0 então
9              se MS_{s,a} existir então
10                 MS_{s,a} ← (1 − d).MS_{s,a} + d.F
11             fim
12             senão
13                 MS_{s,a} ← F
14             fim
15         fim
16         Q(s, a) ← Q(s, a) + α[r + γ. max Q(s', :) − Q(s, a)]
17         s ← s'
18     fim
19 fim
20 Retorna
```

6. Metodologia

Neste trabalho testaremos duas hipóteses sobre o modelo proposto:

- Hipótese H_1: O aprendizado por reforço somático acelera o processo de treinamento dos agentes.
- Hipótese H_2: O aprendizado por reforço somático aumenta a acurácia do agente em execução com alterações no ambiente com relação ao treinamento.

Para testar essas hipóteses, desenvolve-se dois agentes na linguagem Python, capazes de resolver o desafio do *FrozenLake-v1*, da biblioteca *Gymnasium*[1]. O Agente1

[1]Gymnasium Documentation - Farama Foundation - https://gymnasium.farama.org

Tabela 1. Cenários de Teste Executados

Cenário	Dimensão do tabuleiro	Episódios de treino
1	8x8	1000
2	10x10	2000
3	12x12	5000

utiliza o algoritmo clássico do Q-Learning. O Agente2, utiliza o Q-Learning Somático (QLS) proposto.

Como métrica para testar a hipótese H_1, medimos a curva de aprendizado dos agentes, ou seja, quantidade de episódios de treinamento necessários para convergir a resposta dos Agentes (quantidade de passos até o sucesso). Para testar a hipótese H_2, medimos a quantidade de sucessos em cem tentativas de solução dos Agentes em um ambiente levemente diferente do ambiente de treinamento.

Os cenários de teste executados estão descritos na tabela 1.

Todos os treinos são realizados com os mesmos parâmetros do Q-Learning:

- Taxa de aprendizagem (α): 0,8
- Fator de desconto (γ): 0,95
- ϵ: 0,1

Os parâmetros do Q-Learning Somático são:

- Limite de agradabilidade (la): 0,6
- Limite de desagrabilidade (ld): -0,6
- Fator de desconto da memória somática (d): 0,8
- Fator de ajuste (f): 1

O ambiente do *FrozenLake-v1* é definido por um tabuleiro onde o jogador deve deslocar o personagem uma casa por vez, em qualquer uma das quatro direções (direita, esquerda, cima e baixo) até chegar ao presente. Cada casa do tabuleiro simboliza um estado e pode ser gelo rígido ou um buraco de água. Quando atingir o presente, o agente recebe a recompensa no valor 1. Nos demais estados a recompensa é 0. O jogador sempre começa na extremidade esquerda superior do tabuleiro e o presente sempre está na extremidade direita inferior do tabuleiro. O tabuleiro utilizado em cada dimensão é sempre o mesmo para os dois agentes.

Como a recompensa deste ambiente é zero ou um, o fator de ajuste do marcador somático é um (1) pois não existe ajuste de grandeza uma vez que o valor da resposta somática varia de -1 a +1.

No cenário de solução do *FrozenLake-v1*, foram desenvolvidas três funções de emoções sintéticas:

- frustração: retorna um valor somático de -0,7 quando o agente executar uma ação e não mudar de estado
- dor: retorna um valor de -1,0 quando o agente cair em um buraco
- euforia: retorna um valor +1,0 quando o agente encontrar um presente

O conjunto de funções do repositório de Emoções é desenvolvido especificamente para o problema a ser resolvido pelo agente, ou seja, é específico ao ambiente em questão.

Os valores de la e ld são escolhidos levando-se em conta os valores produzidos pelas funções de emoção. Desta forma, os estados onde as ações produzirem "frustração" serão marcados com valores que vão exceder o ld, fazendo com que o agente remova essa ação das possibilidades nesse estado na Avaliação Somática, o mesmo acontecendo com ações que produzirem "dor". Já, quando uma ação produzir "euforia" será marcada com um valor que supera o la, fazendo com que essa ação seja selecionada "somaticamente" na etapa do Sentir quando o agente se encontrar no estado marcado.

Os códigos fonte dos agentes desenvolvidos estão disponíveis em `https://github.com/apcarneiro/bracis_SQL`.

7. Resultados

Uma imagem com o mapa de calor das tabelas Q resultantes do treinamento podem ser observadas na figura 2, bem como o tabuleiro 8x8 utilizado.

O treinamento dos agentes QLS resultam em memórias somáticas com marcações para cada resultante do sentimento, para cada ação, em cada estado. Na figura 3 temos um mapa de calor descrevendo as ações marcadas e qual a intensidade da marca cada estado. As ações em branco (sem marcas) não existem na memória somática. Elas são exibidas nesta figura apenas para criar um paralelo com o tabuleiro utilizado. Sentimentos negativos são marcados em graduações da cor vermelha e positivos em graduações da cor azul.

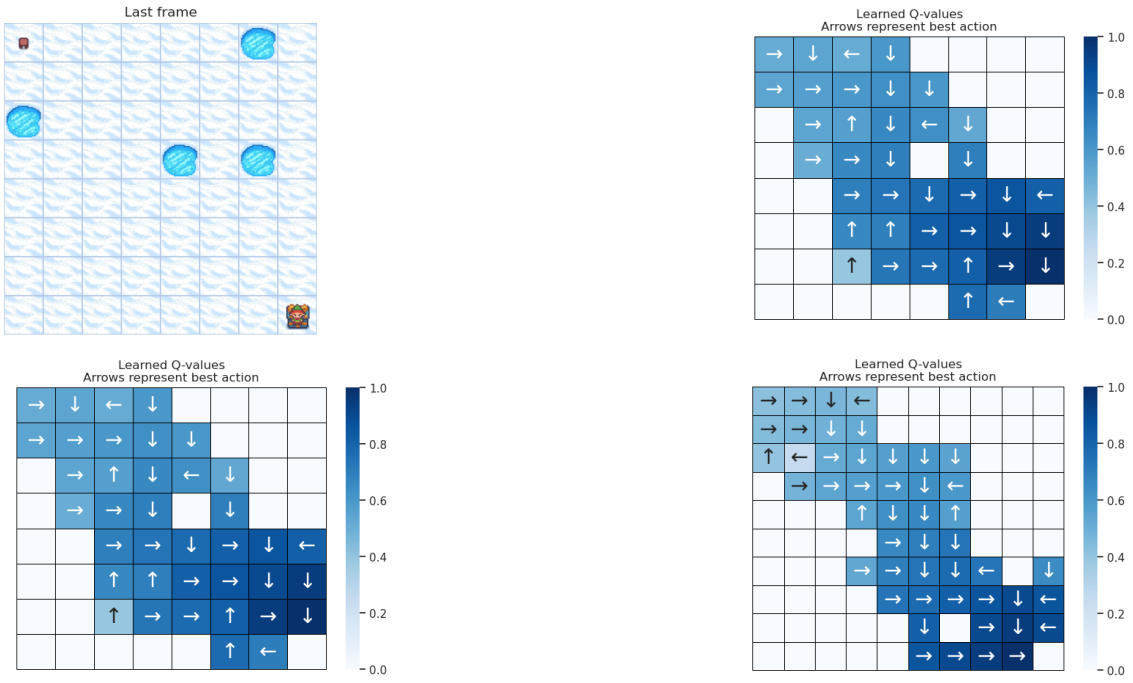

Figura 2. Tabuleiro e mapas de calor das tabelas Q resultantes (fonte: Do Autor)

É possível notar uma estrutura na memória somática (ver figura 3), com marcações negativas nas bordas do tabuleiro, para ações que não tem efeito de mudança de estado (movimento), indicando o funcionamento da emoção sintética *frustração*. Também notamos na memória somática a marcação intensas das ações que levam o agente para os

buracos, indicando o funcionamento da emoção sintética *dor*, e as marcações intensas positivas nas ações que levam o agente para o presente, indicando o funcionamento da emoção sintética *euforia*.

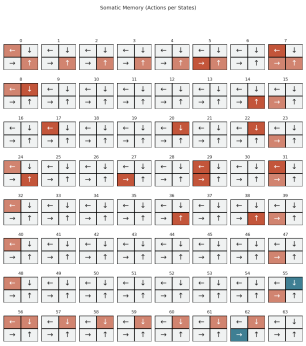

Figura 3. Mapa de calor da memória somática (8x8) (fonte: Do Autor)

Observa-se também os gráficos de acúmulo de recompensas por episódios e média de passos por episódio (figura 4) indicando a curva de aprendizado dos agentes. Para obter esses gráficos os episódios de treinamento foram repetidos 20 vezes em cada um dos agentes e tabuleiro. Nos gráficos de Recompensas Acumuladas por Episódios é exibido, além da linha da progressão média do acúmulo, um sombreado indicando a variação dos valores entre as 20 repetições.

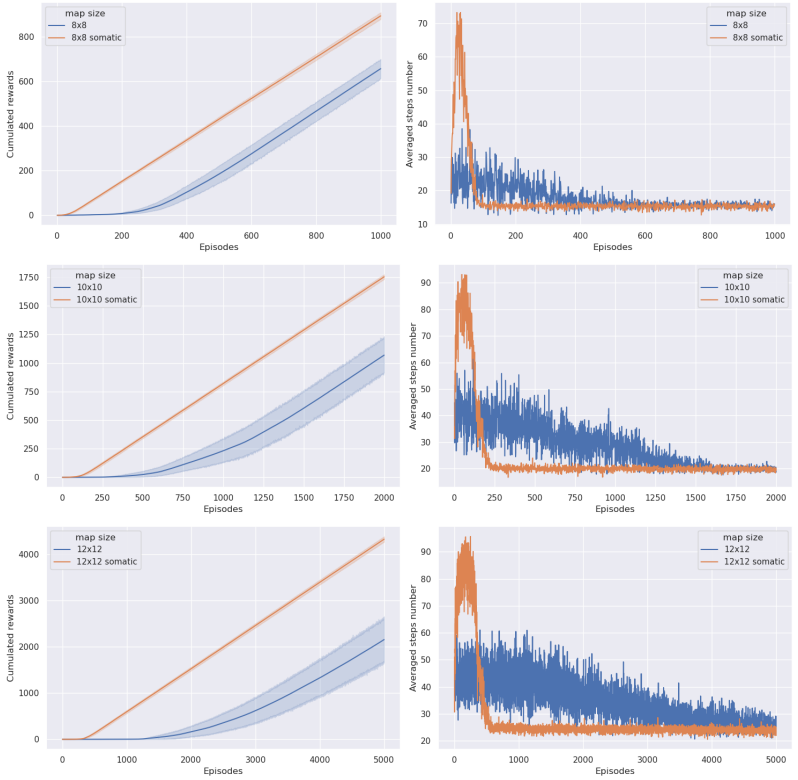

Figura 4. Acúmulo de recompensas e curva de aprendizado (fonte: Do Autor)

A tabela 2 demonstra os valores aproximado de passos e em que episódio do treinamento cada agente convergiu sua resposta, para cada tabuleiro.

Tabela 2. Valor aproximado de passos e episódios para convergência da resposta nos Cenários/Agentes

Cenário	Agente	Episódios	Passos
8x8	QL	400	15
8x8	QLS	50	15
10x10	QL	1500	20
10x10	QLS	200	20
12x12	QL	-	-
12x12	QLS	500	23

Nos testes de acurácia adicionou-se um buraco em cada um dos cenários, posicionando-os no caminho ideal para o presente. Os novos buracos ficaram então nos estados:

- 45 do 8x8 (posição $(y, x) = (5, 5)$)
- 45 do 10x10 (posição $(y, x) = (4, 5)$)
- 90 do 12x12 (posição $(y, x) = (7, 6)$)

O agente clássico não conseguiu resolver o tabuleiro em nenhuma das 100 tentativas, nos cenários 8x8 e 10x10. No cenário 12x12 ele conseguiu resolver em 51 das 100 tentativas.

O agente QLS resolveu o tabuleiro em 99 vezes de 100, nos cenários 8x8 e 10x10, e 98 vezes em 100 no cenário 12x12.

8. Conclusões

Avaliando os gráficos de cada cenário, notamos que os agentes QLS executam uma maior quantidade de passos nos primeiros episódios do treinamento, até que encontre pela primeira vez o presente e propague a recompensa na tabela Q. Isso deve-se ao fato dos marcadores somáticos impedirem que caia nos buracos constantemente (como acontece com o agente clássico) fazendo com que ele fique tentando caminhos por mais passos em cada episódio.

Uma vez encontrado pela primeira vez o presente, o agente QLS converge a resposta muito rapidamente. Isto é notado nos gráficos das figura 4. Na tabela 2 temos o agente QLS convergindo para a resposta oito vezes mais rápido no cenário 8x8 e 7,5 vezes mais rápido no cenário 10x10. No cenário 12x12, apesar de executarmos o treinamento por 5000 episódios, o agente clássico não convergiu. Porém, o agente QLS convergiu em 500 episódios, para uma resposta média muito próxima do ideal que são 22 passos.

Outra diferença notável está na acurácia do agente QLS quando existe uma pequena alteração no tabuleiro. Ele mostrou-se capaz de se adaptar rapidamente à alteração, cometendo apenas um erro nos cenários 8x8 e 10x10 e dois erros no cenário 12x12, em cem tentativas enquanto o agente clássico não consegue resolver ou resolve de maneira muito mediana dependendo do cenário.

Esses resultados demonstram as hipóteses H_1 e H_2 como verdadeiras para o Q-Learning Somático, no ambiente do Frozenlake, indicando que o modelo mostra-se promissor quanto à melhora dos algoritmos de RL.

É importante destacar que, outras técnicas de otimização dos algoritmos de aprendizado por reforço como o uso de heurísticas para escolha das ações

[Bianchi et al. 2009] ou modulações das recompensas recebidas [Eschmann 2021] podem acelerar também o processo de aprendizado, porém, o modelo de aprendizado por reforço somático cria uma memória somática de estados e ações que continuam sendo alteradas ou incrementadas na "vida" do agente, mesmo após a fase de treinamento, tornando ele mais resiliente à alterações no ambiente. Além disso, o conjunto de funções emoção pode ser alterado, adicionando, removendo ou alterando funções já existentes, de maneira a alterar também o comportamento do agente perante o ambiente, sem a necessidade de um novo treinamento. Similar ao que acontece com seres vivos que vão desenvolvendo suas características emotivas durante a vida.

Trabalhos futuros serão desenvolvidos demonstrando o uso deste modelo com algoritmos mais atuais e ambientes mais complexos utilizados no aprendizado por reforço, além de ensaios para demonstrar as mudanças no comportamento do agente perante alterações dos limites ld, la e o fator de desconto da memória somática d.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Bianchi, R., Ros, R., and Mantaras, R. (2009). Improving reinforcement learning by using case based heuristics. pages 75–89.

Cabrera-Paniagua, D., Flores, D., Rubilar-Torrealba, R., and Cubillos, C. (2023). Bio-inspired artificial somatic index for reflecting the travel experience of passenger agents under a flexible transportation scenario. *Scientific Reports*, 13(1).

Cominelli, L., Mazzei, D., and De Rossi, D. E. (2018). Seai: Social emotional artificial intelligence based on damasio's theory of mind. *Frontiers in Robotics and AI*, 5.

Damásio, A. (1994). *O Erro de Descartes*. Companhia das Letras.

Eschmann, J. (2021). *Reward Function Design in Reinforcement Learning*, pages 25–33. Springer International Publishing.

Hoefinghoff, J. and Pauli, J. (2013). Reversal learning based on somatic markers. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 498–504.

Maçãs, M., Ventura, R., Custódio, L., and Pinto-Ferreira, C. (2001). DARE: an emotion-based agent architecture. In Russell, I. and Kolen, J. F., editors, *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference, May 21-23, 2001, Key West, Florida, USA*, pages 150–154. AAAI Press.

Pimentel, C. F. and Cravo, M. R. (2009). "don't think too much!" — artificial somatic markers for action selection. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*.

Russell, S. and Norvig, P. (2021). *Artificial Intelligence - A Modern Approach - Fourth Edition*. Person Education Limited.

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction - Second Edition*. The MIT Press.