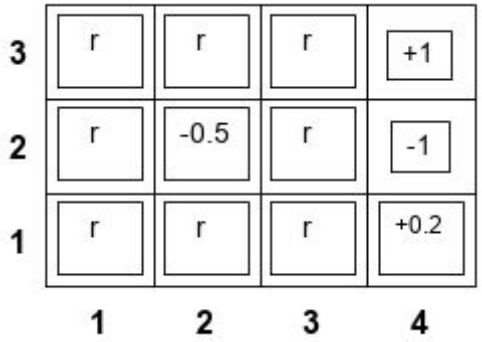
Agentes Cognitivos e Adaptativos 2020.1
Alunos: Diana Marcela e Vinícius Andrade

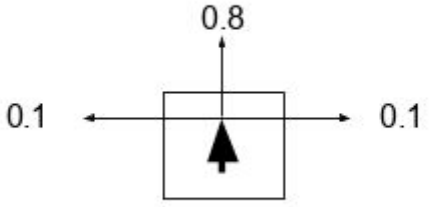
<u>Exercício Decisões Sequenciais</u>

Implemente o algoritmo de MDP com as equações de Bellman e aplique no exercício em anexo. O exercício é similar ao ambiente 4 x 3 visto nas aulas. No entanto, a posição não é mais um obstáculo e sim uma casa onda há um recompensa igual a -0.5. Além disso, existe uma casa com recompensa positiva igual +0.2.

- **Exemplo: Ambiente 4x3**
  - Recompensas R(s):
  - Dois estados finais R(s) = +1 ou R(s) = -1
  - R(s) = -0.5 na posição (2,2)
  - R(s) = +0.2 na posição (4,1)
  - R(s) = r (valor fixo) para todos estados não terminais



- Implemente o algoritmo de MDP com as equações de Bellman e retorna a política retorna para as seguintes situações:
  - r = -0.4
  - r = -0.04
  - r = -0.0004



- Comente as diferenças entre as políticas retornadas
  - As políticas observadas:

- **r = -0.4**

Como apresentado a imagem abaixo com uma recompensa muito baixa chegando a se tornar um custo, o algoritmo procura sempre ir na direção do estado que maximize sua recompensa, e como o estado (1,4) possui uma recompensa ativa de **0.2** as políticas tende a levar o caminho para a mesma com exceção da terceira linha que tende a levar para o estado terminal de utilidade 1 como demonstrado na **Figura 1**.

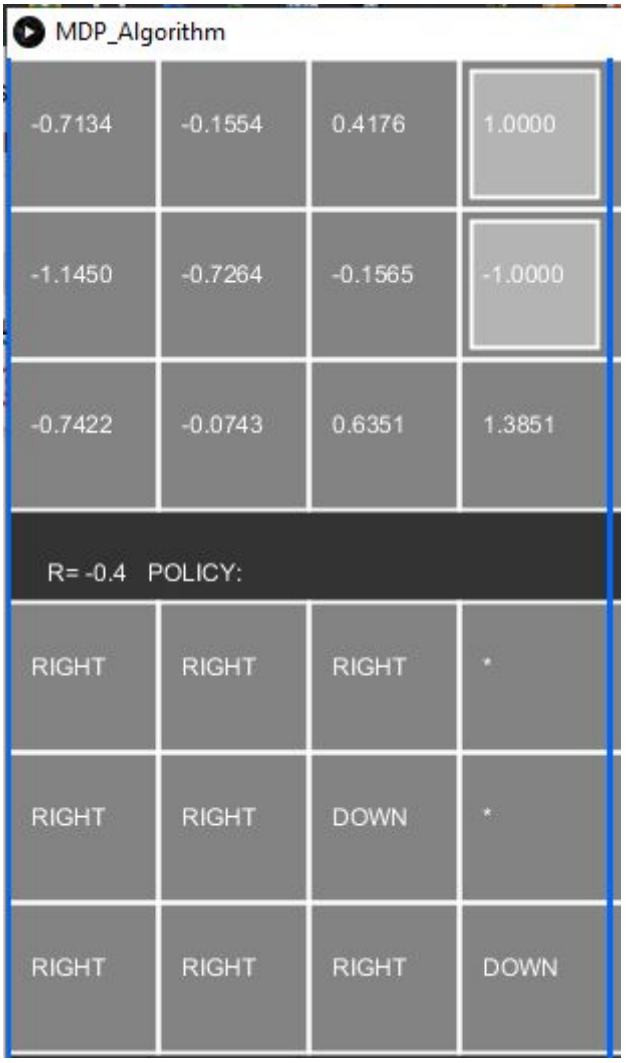


Figura 1

Curiosamente se iniciarmos pelo estado (1,4) ele tenderá a ficar nesse estado indo sempre para baixo e voltando, já o estado (1,3) o redireciona para o estado (1,4). E isto é algo que repete em todos os ciclos abaixo que possuem recompensas menores.

Porém, o mesmo não é visto quando a recompensa do estado (1,4) é negativa, no qual ele sempre tenta fugir desse estado como na **Figura 2**.

Figura 2

Já se a recompensa geral é de **-0.6** for menor que a do estado (1,4), de **-0.2,** ao chegar nesse estado o agente procura terminar logo com o caminho e se mata no estado terminal de (2,4), pois o custo para qualquer outro estado vizinho seria maior que o do terminal negativo.

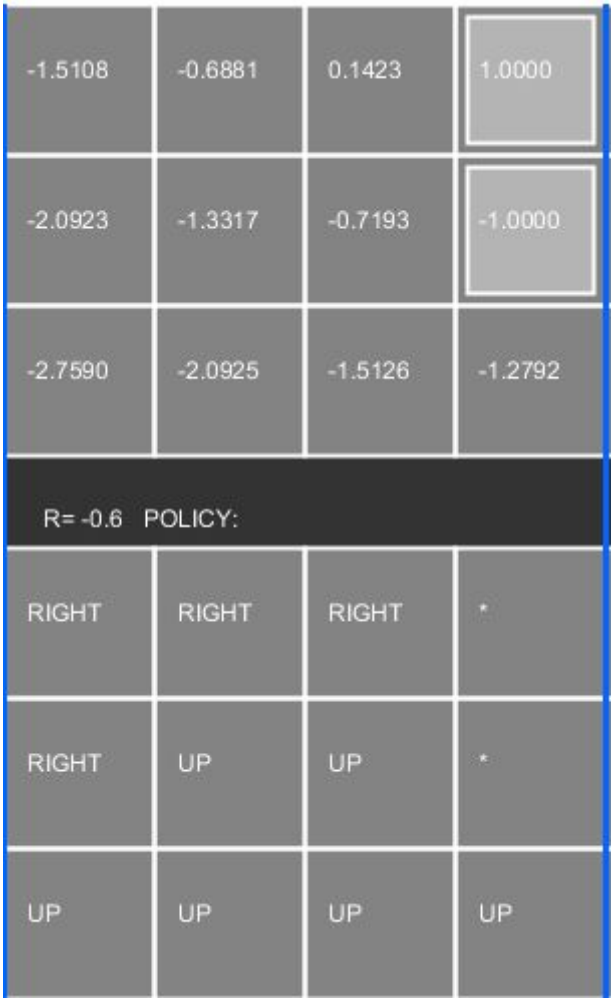


Figura 3

- **r = -0.04**

Já para a recompensa de **-0.04** as primeira interações produzem políticas que tendem ao caminho que leva para o estado terminal (3,4) que possui a utilidade positiva de **1** demonstrado na **Figura 4**, porém após algumas dezenas interações sem muita alteração, à política muda começa a encaminhar para o estado (1,4) que possui uma recompensa de **0.2** e eventualmente se torna superior à utilidade do estado terminal apresentado da **Figura 5**.

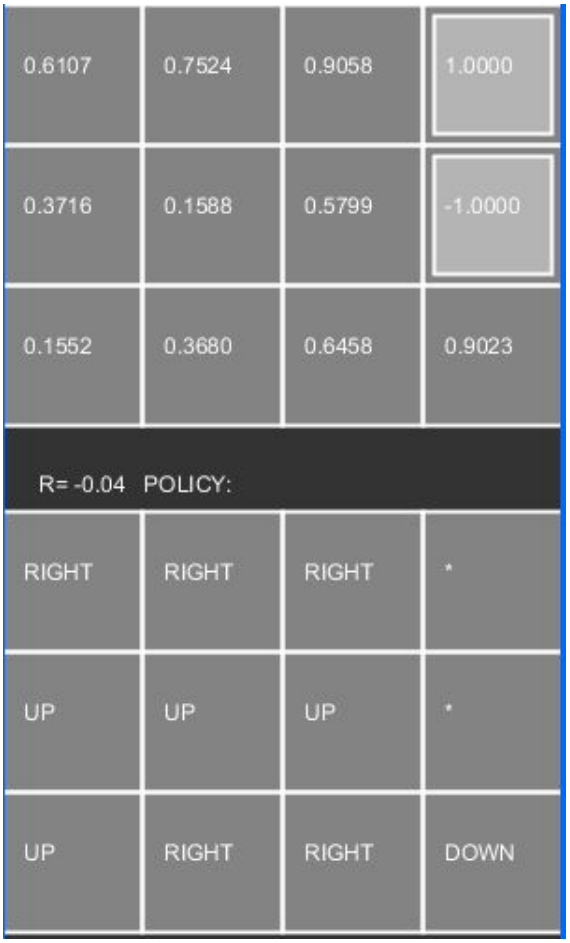


Figura 4: 10 interações

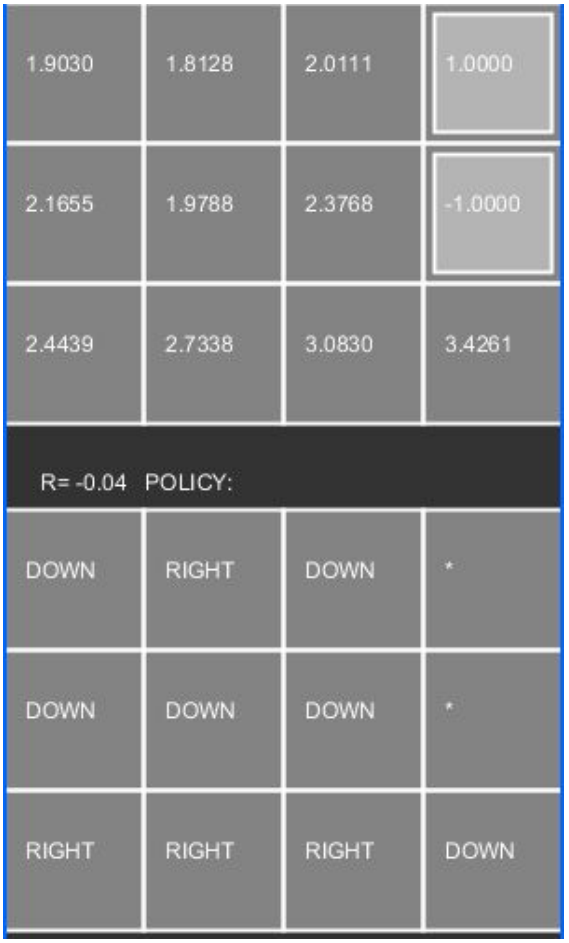


Figura 5: 23 interações

- **r = -0.0004**

De modo geral, a recompensa **-0.0004** produz política que leva a encontrar o caminho do estado terminal positivo (3,4), de valor **1**, mais rápido que os demais por ter um custo muito baixo. Apesar de, estando nas posições (1,2) e (1,3), o agente tende a escolher o caminho que leva ao estado (1,4) com a recompensa **0.2**, como pode ser visualizado na **Figura 6.**

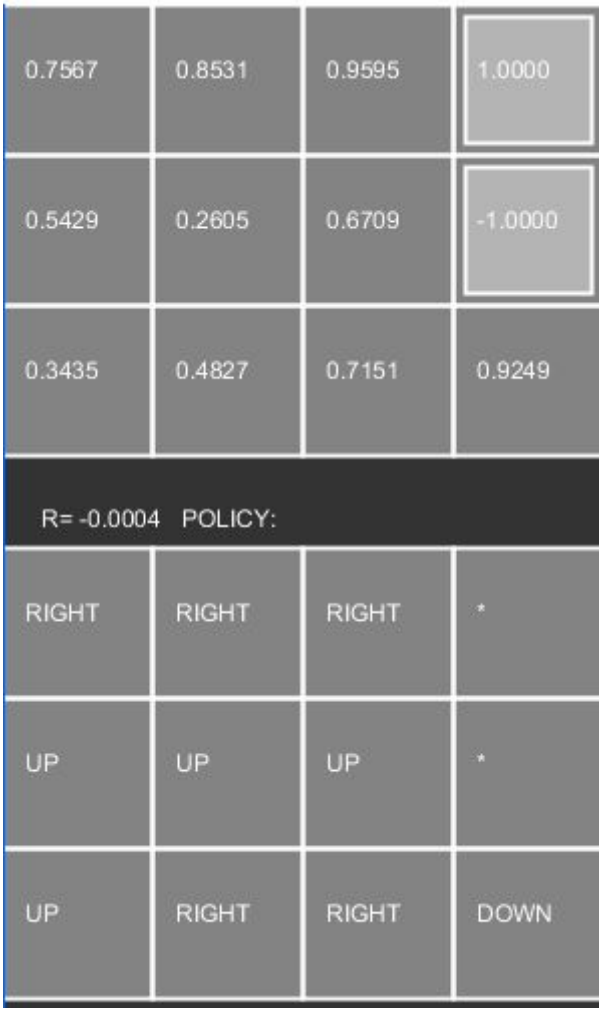

0.7567
0.8531
0.9595
1.0000
0.5429
0.2605
0.6709
-1.0000
0.3435
0.4827
0.7151
0.9249
R= -0.0004 POLICY:
RIGHT
RIGHT
RIGHT
*
UP
UP
UP
*
UP
RIGHT
RIGHT
DOWN

Figura 6

Neste caso, sendo a recompensa geral de **-0.6**, observa-se que o agente tem uma maior tendência a escolher caminhos que o levem a recompensa **-0.2**, na posição (1,4), com exceção se o início se der na posição (3,1), que leva o agente direto ao estado terminal positivo de posição (3,4) e valor **1**.

Figura 7