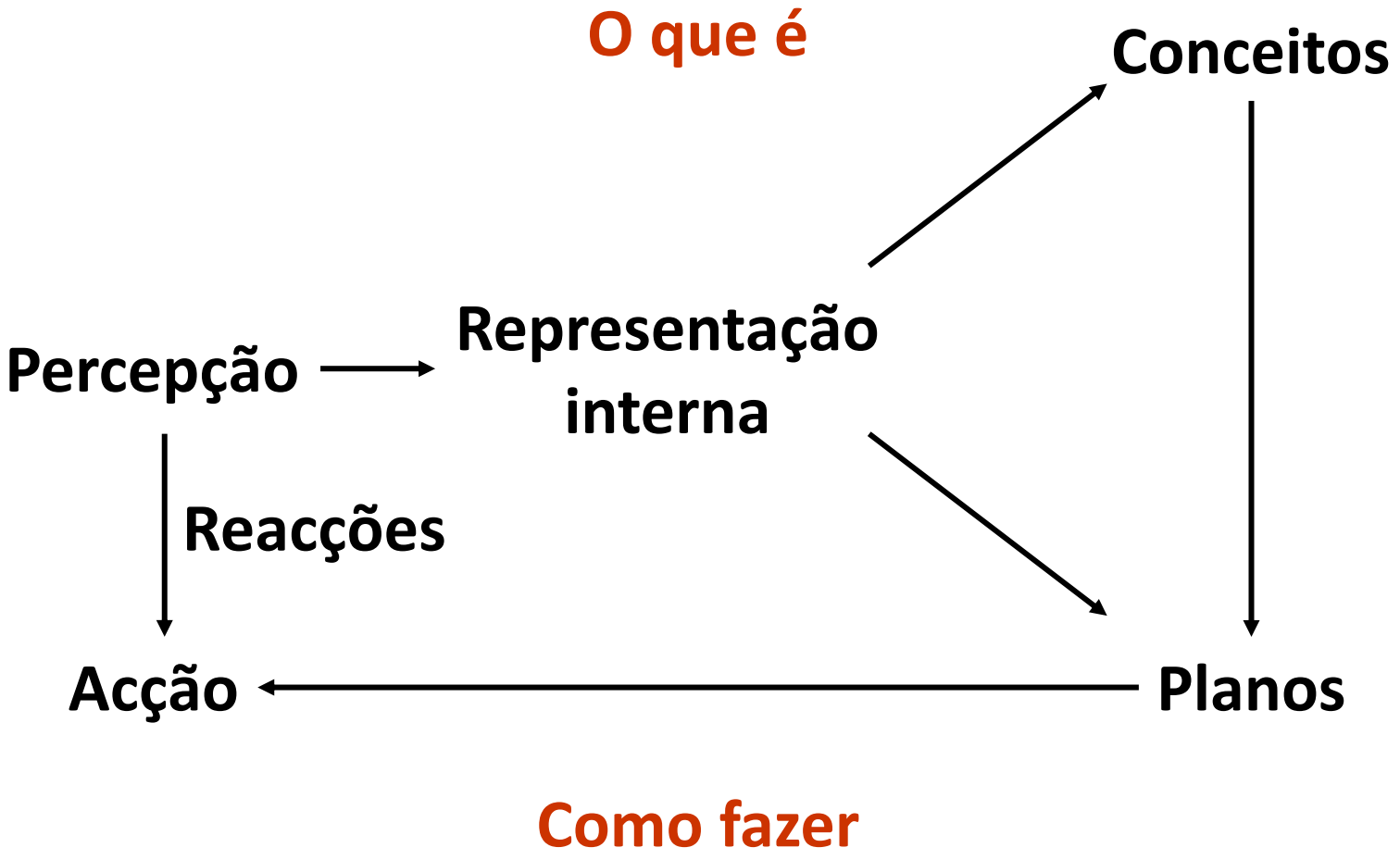
# APRENDIZAGEM POR REFORÇO

Luís Morgado

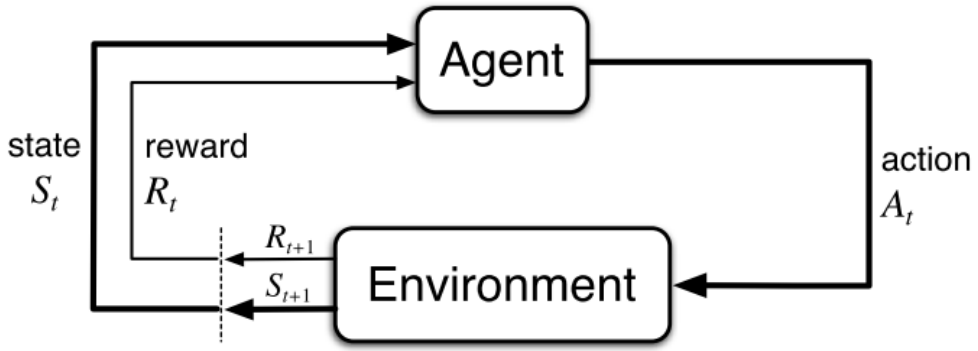
ISEL-ADEETC

# APRENDIZAGEM COMPORTAMENTAL



# APRENDIZAGEM COMPORTAMENTAL

- Aprendizagem de **comportamentos**
  - O que fazer
  - Relação entre situações e acções
- **Aprendizagem por reforço**




[Sutton & Barto, 2020]

# APRENDIZAGEM POR REFORÇO

- Aprendizagem de **comportamentos (padrões de acção)**
  - O que fazer
  - Relação entre situações e acções
- Aprendizagem a partir da **interacção** com o ambiente
- Aprendizagem em **continuidade**

# APRENDIZAGEM POR REFORÇO

Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond. (Thorndike, 1911, p. 244)

Thorndike, E. L. (1911). *Animal Intelligence.* Hafner, Darien, CT.

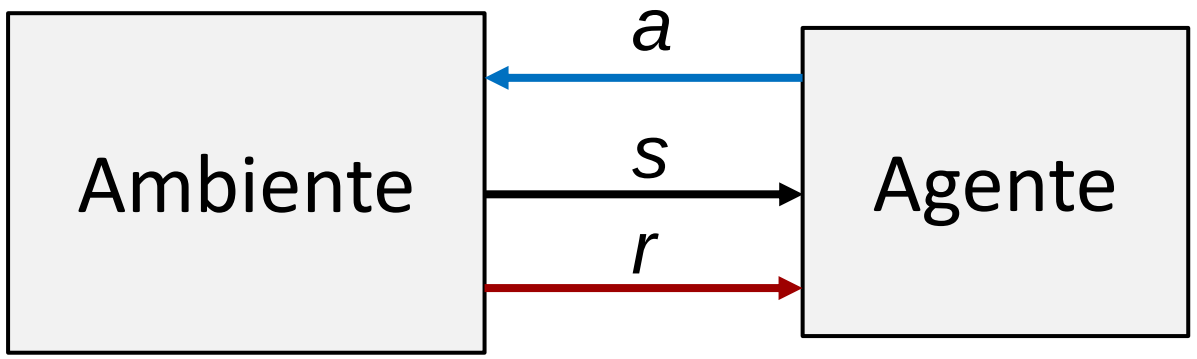
Thorndike called this the "Law of Effect" because it describes the effect of reinforcing events on the tendency to select actions.

*Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal.*

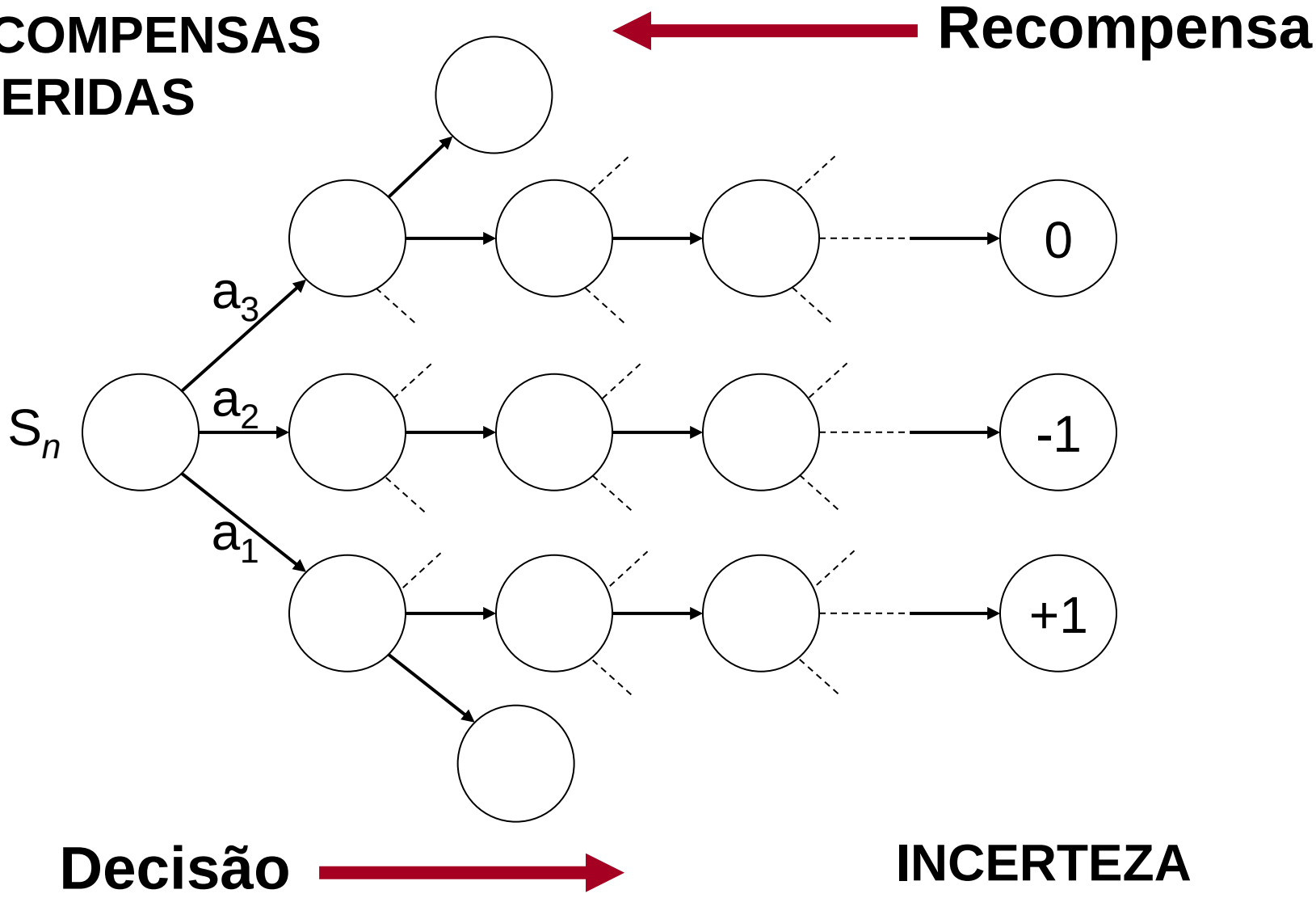
[Sutton & Barto, 2020]

# APRENDIZAGEM POR REFORÇO

- Aprendizagem a partir da **interacção** com o ambiente
  - **Acção**
  - **Estado**
  - **Reforço**
    - Ganho / perda



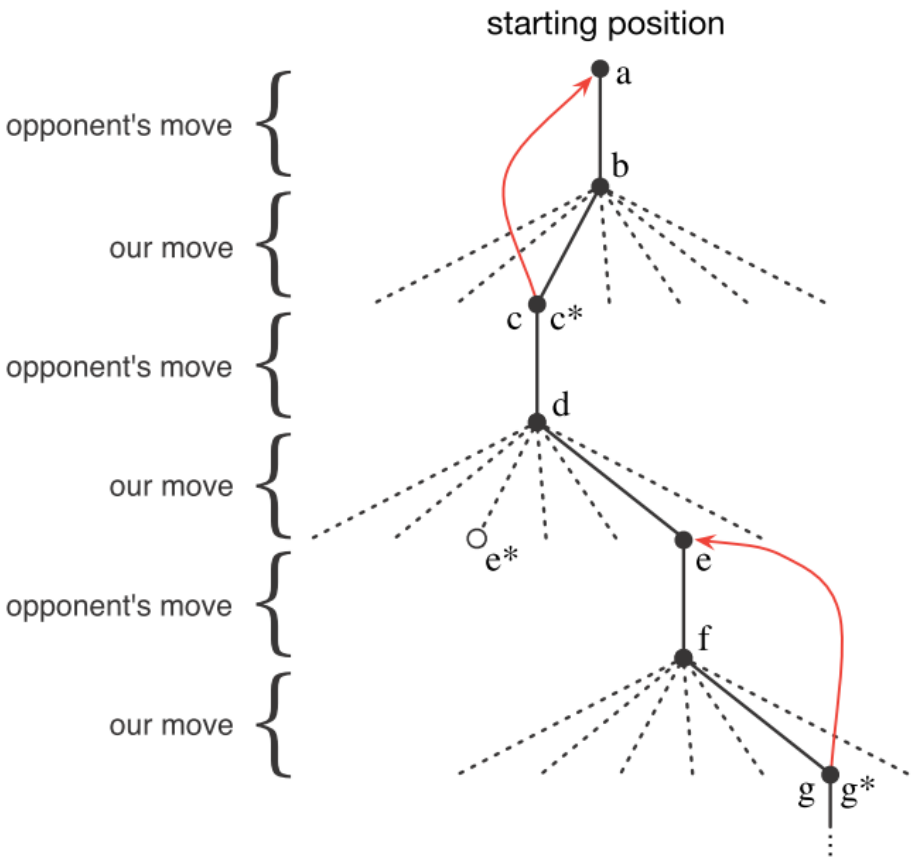
# O PROBLEMA DA DECISÃO SEQUENCIAL



# APRENDIZAGEM POR REFORÇO

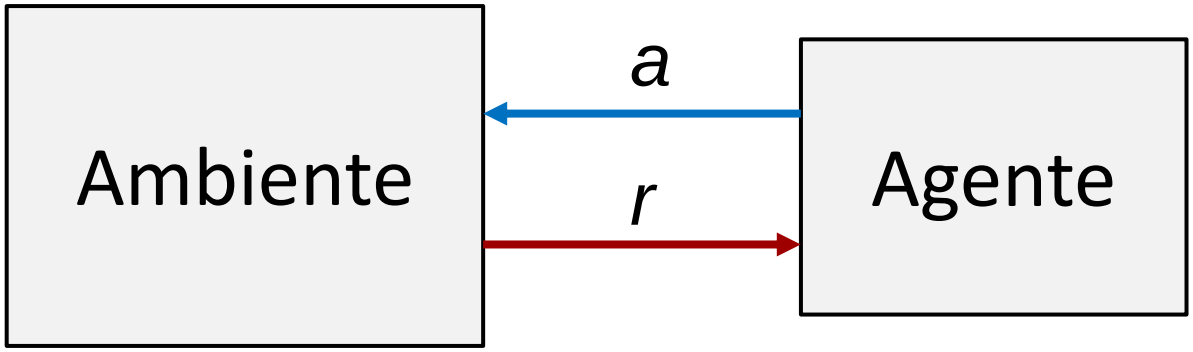
**Exemplo**: Aprender a jogar um jogo

| | | |
|---|---|---|
| X | O | O |
| O | X | X |
| | | X |



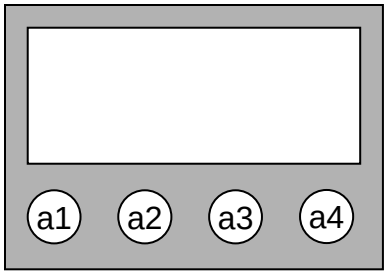

[Sutton & Barto, 2020]

# APRENDIZAGEM DE VALOR DE ACÇÃO



# APRENDIZAGEM DE VALOR DE ACÇÃO

- Exemplo: escolha repetida de diferentes acções
- Por cada acção é obtida uma recompensa
- Resultado depende só da acção escolhida
- **Motivação**
  - **Maximizar a recompensa de longo prazo**



# APRENDIZAGEM DE VALOR DE ACÇÃO

- Como determinar o valor ($Q$) de cada acção?

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t}$$

- Valor médio para uma acção $k$ após $n$ tentativas

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

[Sutton & Barto, 2020]

# FORMA INCREMENTAL

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}$$

$$NewEstimate \leftarrow OldEstimate \; + \; StepSize \Big[Target - OldEstimate\Big]$$

$\Big[Target - OldEstimate\Big]$ is an *error* in the estimate

[Sutton & Barto, 2020]

# APRENDIZAGEM DE VALOR DE ACÇÃO

**A simple bandit algorithm**

Initialize, for $a = 1$ to $k$:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \operatorname{arg\,max}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$

[Sutton & Barto, 2020]

# APRENDIZAGEM DE VALOR DE ACÇÃO

- **Problemas não estacionários?**

- Estimação por acumulação não linear
  - Por exemplo, exponencialmente amortecida

$$Q_{n+1} = Q_n + \alpha\left[R_n - Q_n\right]$$

$\alpha \in$ [0,1] - Factor de aprendizagem

# PROBLEMAS NÃO-ESTACIONÁRIOS

$$Q_{n+1} \doteq Q_n + \alpha\Big[R_n - Q_n\Big] \qquad \text{step-size parameter } \alpha \in (0, 1]$$

$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha\Big[R_n - Q_n\Big] \\
&= \alpha R_n + (1-\alpha)Q_n \\
&= \alpha R_n + (1-\alpha)\left[\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right] \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2\alpha R_{n-2} + \\
&\qquad \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n Q_1 \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i.
\end{aligned}$$

*exponential recency-weighted average*

$\alpha_n(a) = \frac{1}{n}$ results in the sample-average method

[Sutton & Barto, 2020]

# EXPLORAÇÃO vs. APROVEITAMENTO

- Duas vertentes da aprendizagem:
  - **EXPLORAÇÃO** (*explore*)
    - Tem por objectivo explorar todos os estados possíveis, tentando todas as acções
    - Obtenção de experiência
  - **APROVEITAMENTO** (*exploit*)
    - Utiliza o conhecimento resultante da aprendizagem para obter o máximo de recompensa
    - Restringe-se às acções que se conhece serem favoráveis

# DILEMA EXPLORAR / APROVEITAR (*EXPLORE / EXPLOIT*)

- **Quando é que se aprendeu o suficiente para começar a aplicar o que se aprendeu?**

- **Exploração** (*Exploration*)
  - Escolher uma acção que permita explorar o mundo para melhorar a aprendizagem

- **Aproveitamento** (*Exploitation*)
  - Escolher a acção que leva à melhor recompensa de acordo com a aprendizagem
  - Acção Sôfrega (*Greedy*)

# REFERÊNCIAS


[Sutton & Barto, 2020]
R. Sutton, A. Barto, “Reinforcement Learning: An Introduction”, 2nd Edition, MIT Press, 2020

[Poole & Mackworth, 2010]
D. Poole, A. Mackworth, Artificial Intelligence: Foundations of Computational Agents, Cambridge University Press, 2010

[Barnard, 2003]
C. Barnard, “Animal Behaviour: Mechanism, Development, Ecology and Evolution”, Prentice Hall, 2003

[Koppula, 2020]
R. Koppula, “Exploration vs. Exploitation In Reinforcement Learning”, *https://www.manifold.ai/exploration-vs-exploitation-in-reinforcement-learning*, 2020