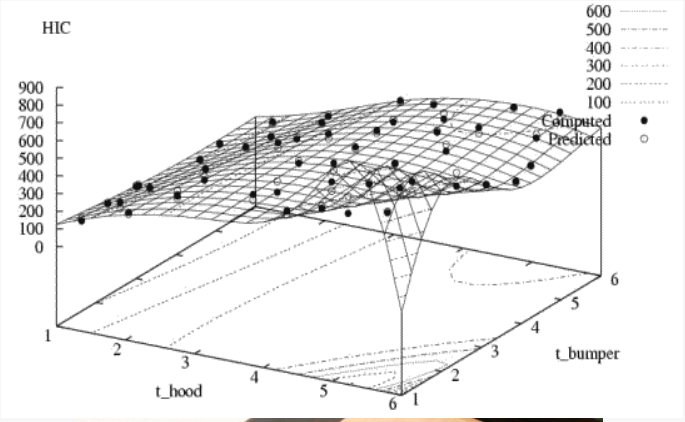
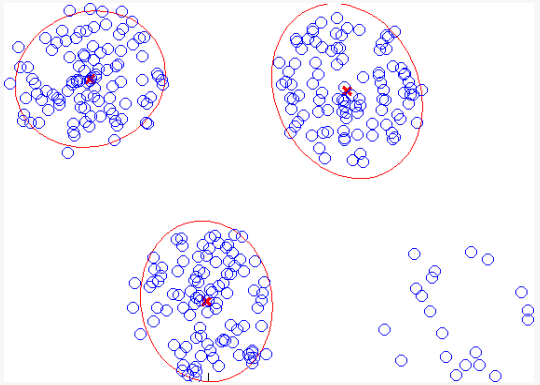
Megerősítéses tanulás

A tanulás alapvető típusai

- Felügyelt
  - Az adat: bemenet-kimenet párok halmaza
  - A cél: függvényapproximáció, klasszifikáció
- Megerősítéses
  - Az adat: állapotmegfigyelések és jutalmak
  - A cél: optimális stratégia a jutalom maximalizálására
- Nem felügyelt, reprezentációs
  - Az adat: bemenetek halmaza
  - A cél: az adat optimális reprezentációjának megtalálása / magyarázó modell felírása
- Egymásba ágyazások

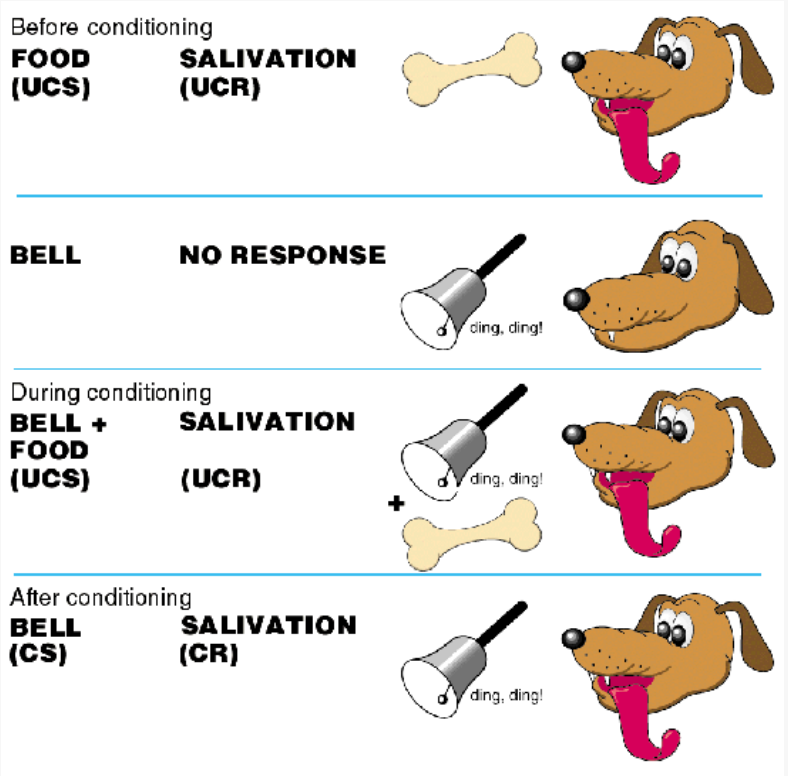




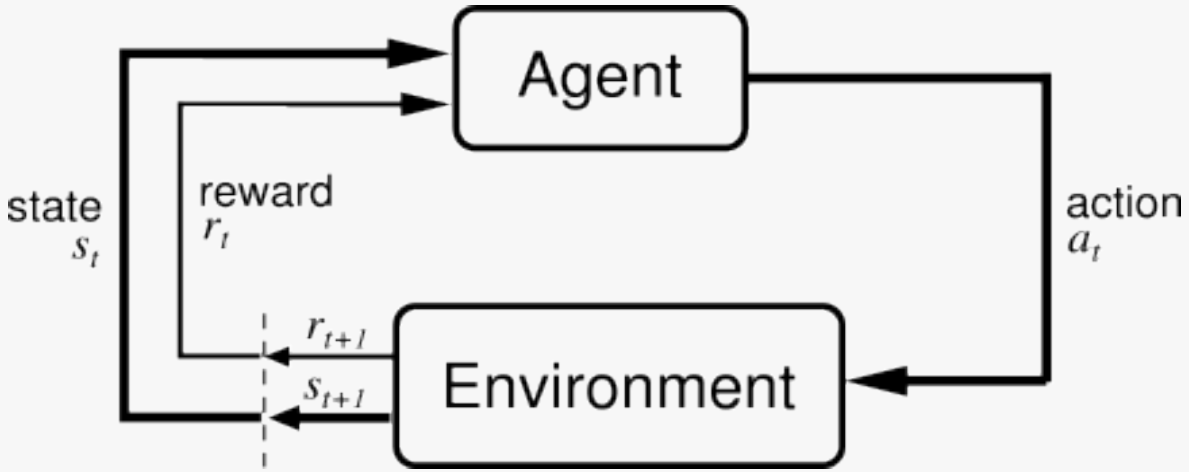
Klasszikus kondicionálás

- Rescorla-Wagner szabály
- A feltételes és a feltétel nélküli stimulusok közti asszociáció erősségét tanulja
- Jó heurisztika, sok problémája van

$$\Delta V_A = \alpha_A \beta (\lambda - V_{AX})$$
$$\Delta V_X = \alpha_X \beta (\lambda - V_{AX})$$
$$V_{AX} = V_A + V_X$$



Az ágens-környezet modell



- Ágens: a tanulórendszer
- Környezet: minden, amit nem belsőleg állít elő
- Bemenet: aktuális állapot, jutalom
- Kimenet: cselekvés

A jutalom hosszútávú maximalizációja

- Összesített jutalom: valamilyen becslést kell alkalmazni a jövőre nézve
- Felfedezés vagy kihasználás dilemmája
- Különbség a felügyelt tanulástól: egyiknél azt tudjuk, hogy a képtérben milyen messze vagyunk az optimálistól, a másiknál arra kapunk egy becslést, hogy a paramétertérben mennyire térünk el az optimálistól

Az állapottér reprezentációja

- Állapotváltozók
- Értékfüggvény: állapot, cselekvés
- Táblázat
- Függvényközelítők
- Neurális hálózat
- Értékfüggvény reprezentációjának tanulása: lépésenként, supervised módon, hiba konstruálásával
- Ha az állapottér struktúráját is tanulni kell: representational learning

Tanulási szabályok

- Modellalapú TD-learning
  - Markov-lánc átmenetvalószínűségei (MDP)
  - Állapotokhoz rendelt értékek tanulása

$$V_{t+1}(s_t) = V_t(s_t) + \alpha[r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)]$$

- Modellfüggetlen: Q-learning
  - Állapot-cselekvés párokhoz rendelt értékek tanulása

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[r_{t+1} + \gamma \, max_a Q_t(x_{t+1}, a) - Q_t(s_t, a_t)]$$

- Visszaterjesztés korábbi állapotokra: λ felejtési paraméterrel súlyozva módosítjuk a korábban tapasztalt állapotok vagy állpot-cselekvés párok értékeit is az aktuális deltával
- On-policy szabályok (SARSA): nem a lehető legjobb, hanem mindig az aktuálisan választott cselekvés alapján tanulunk

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[r_{t+1} + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(s_t, a_t)]$$

TD neurális reprezentációval

- Prediction error felhasználása a tanuláshoz
- Az állapotérték frissítése neurális reprezentációban:

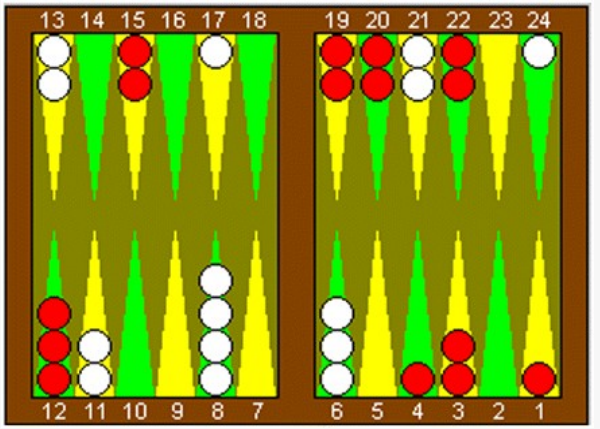
$$w(\tau) \rightarrow w(\tau) + \epsilon\, \delta(t)\, u(t-\tau) \qquad \delta(t) = \sum_{\tau} r(t+\tau) - v(t)$$

- A prediction error kiszámítása
  - A teljes jövőbeli jutalom kellene hozzá
  - Egylépéses lokális közelítést alkalmazunk

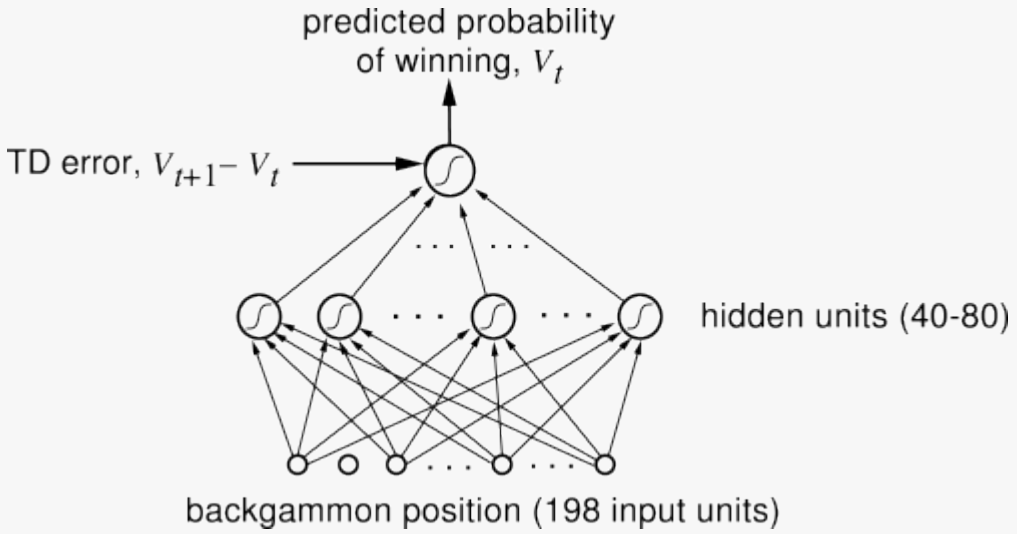
$$\sum_{\tau} r(t+\tau) - v(t) \approx r(t) + v(t+1)$$

  - Ha a környezet megfigyelhető, akkor az optimális stratégiához konvergál
- A hibát visszaterjeszthetjük a korábbi állapotokra is (hasonlóan a backpropagation algoritmushoz)
- Akciókiválasztás: exploration vs. exploitation

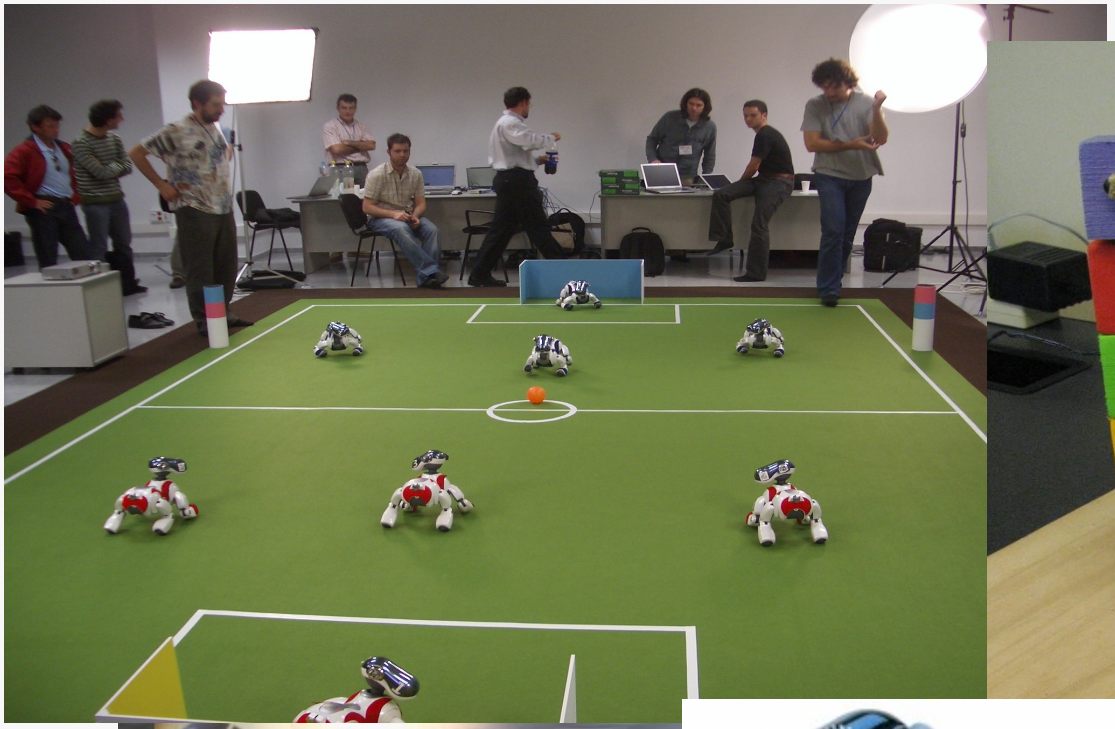
TD tanulás neurális hálózattal

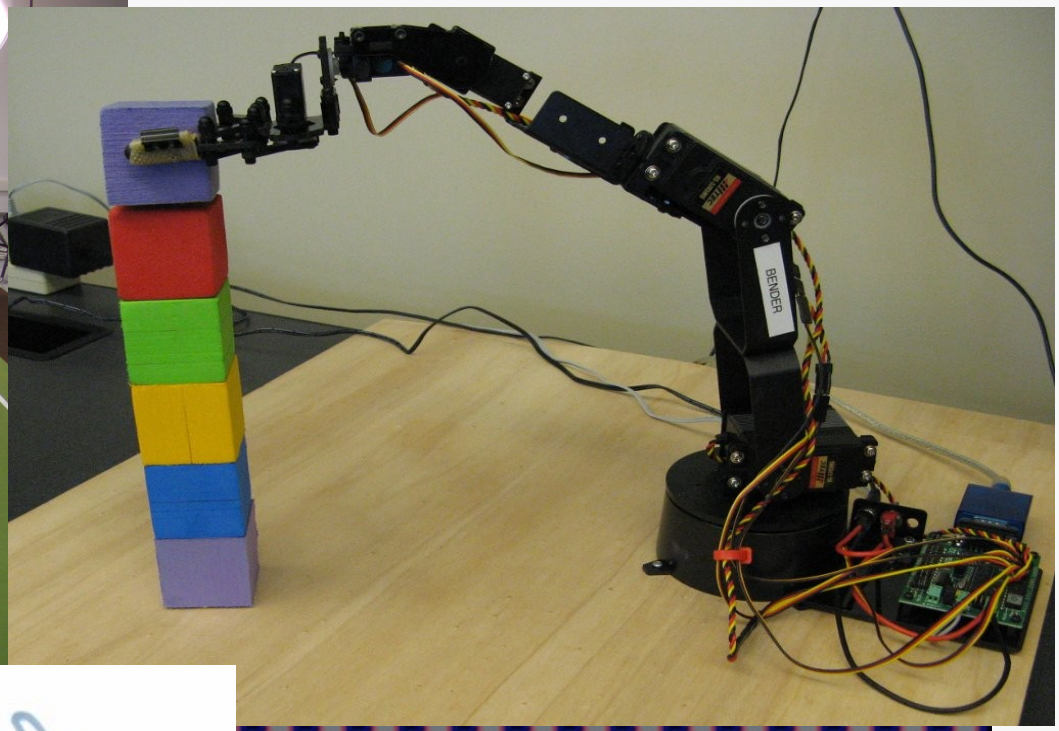


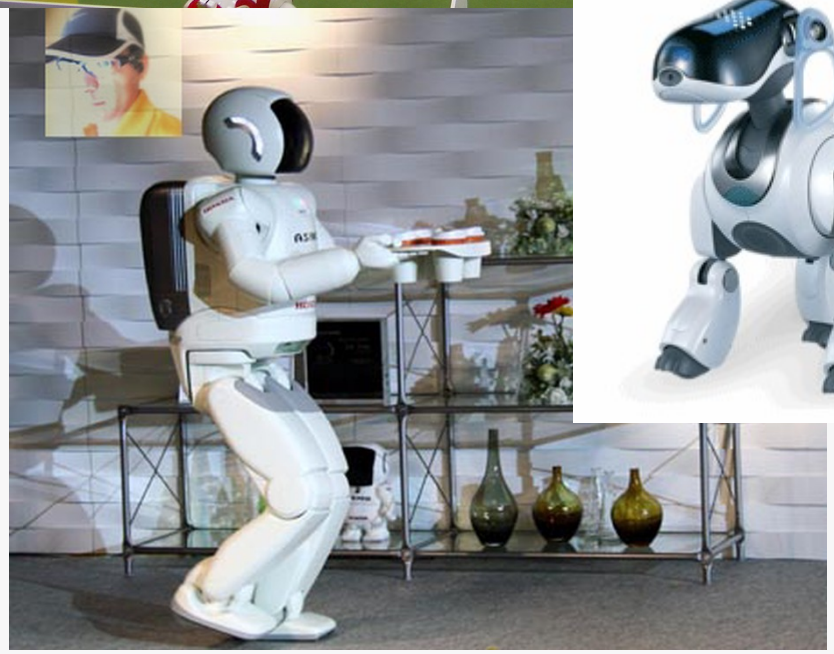
- Gerald Tesauro: TD-Gammon
  - Előrecsatolt hálózat
  - Bemenet: a lehetséges lépések nyomán elért állapotok
  - Kimenet: állapotérték (nyerési valség)
- Minden lépésben meg kell határozni a hálózat kimeneti hibáját
  - Reward signal alapján
- Eredmény: a legjobb emberi játékosokkal összemérhető

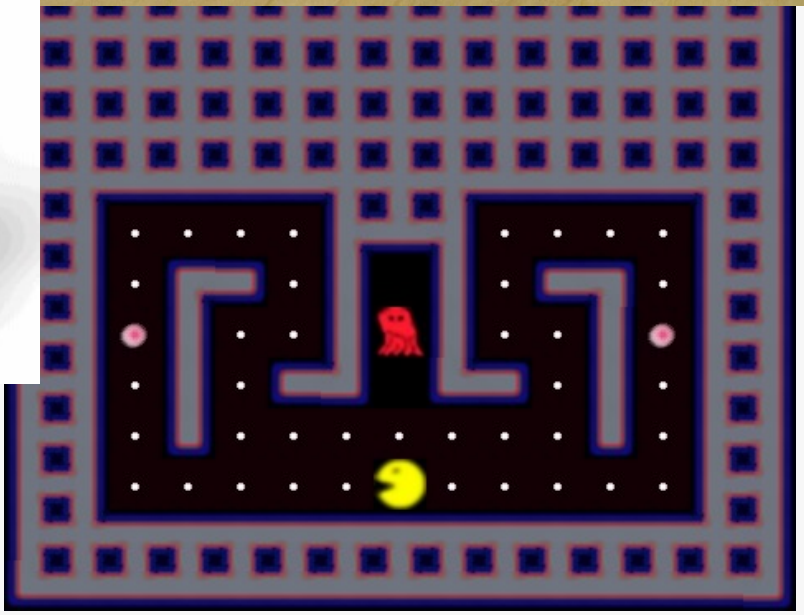


Gépi példák









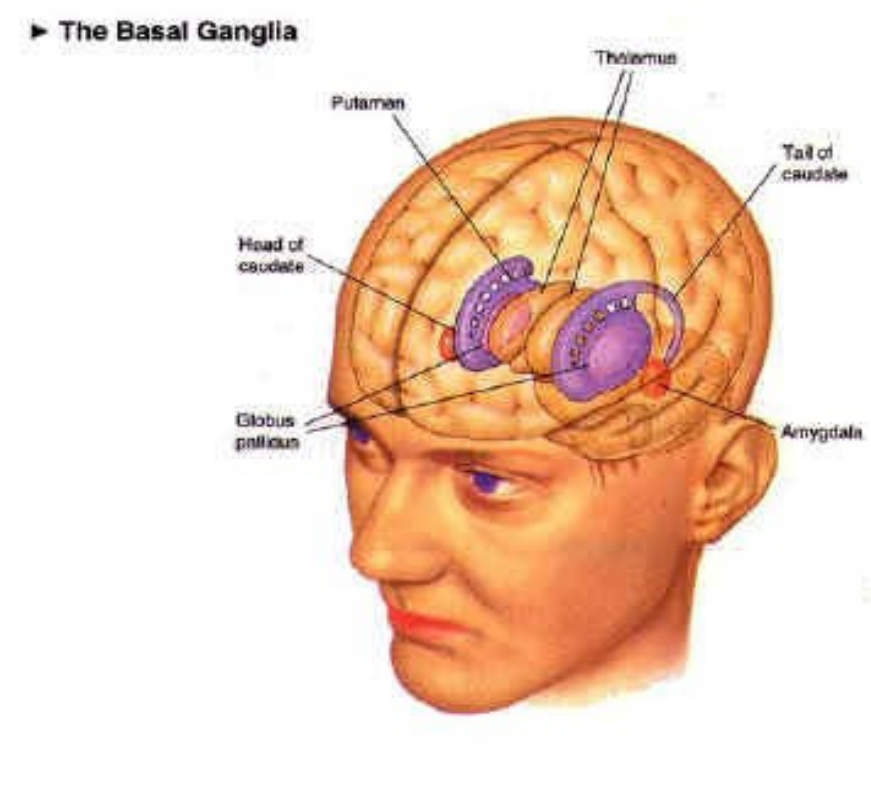
Megerősítéses tanulás az idegrendszerben

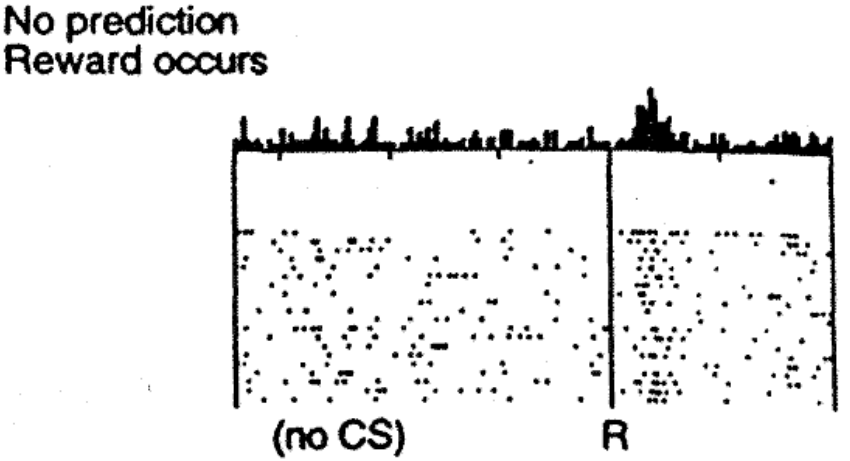
- Jutalomszignál: dopamin a bazális ganglionokban
- Modellalapú és modell nélküli megerősítéses tanulórendszere
- Az idegrendszer valószínűleg kombinálja a kettőt, bizonyossággal súlyozva

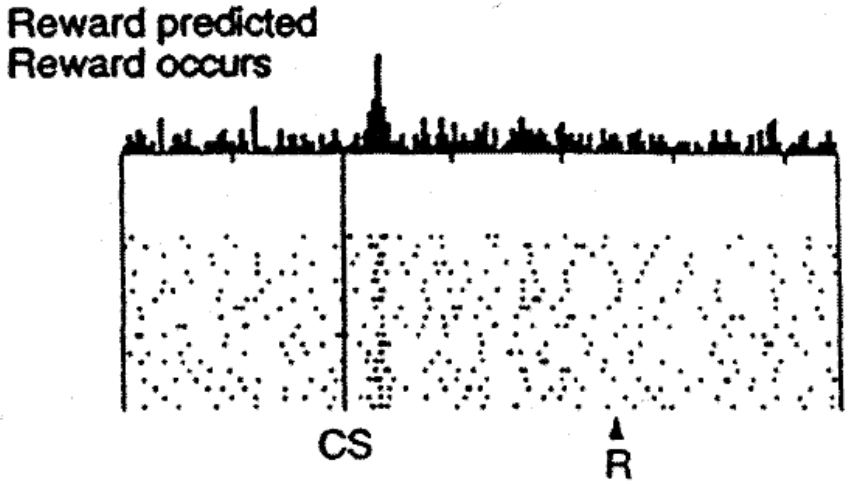
The effect of reward in dopaminerg cell of basal ganglia

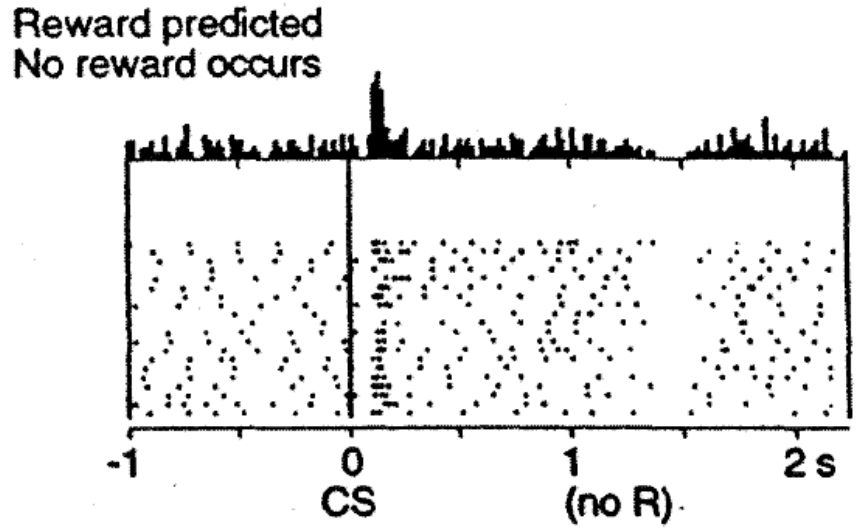
An interpretation:

Dopamine cells signals the difference between the expected and received reward.

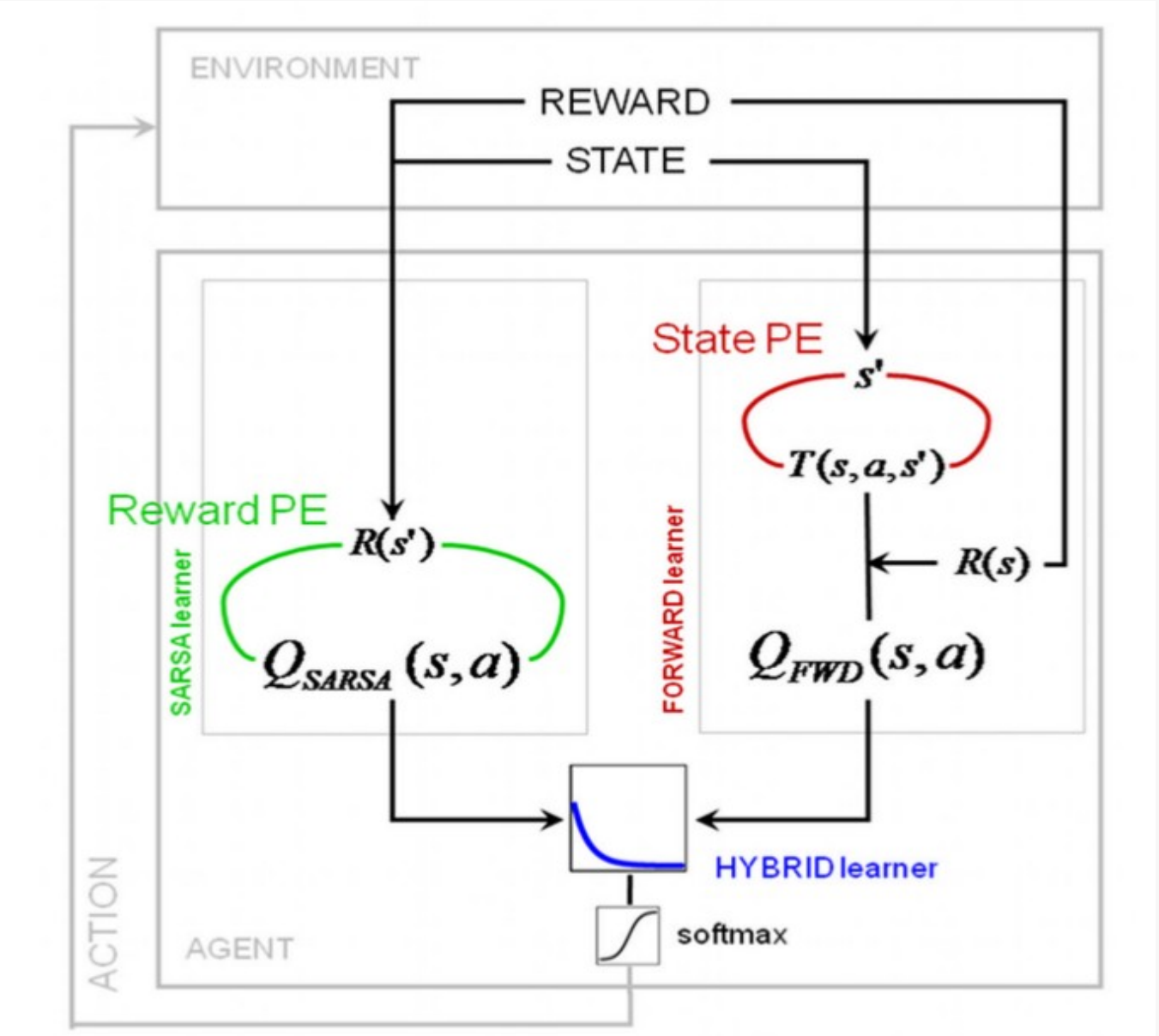








- Glascher, Daw Dayan, O'Doherty, *Neuron,* 2010.
- Modellalapú és modell nélküli algoritmusok predikciójával
- való korreláció



- fMRI eredmények