*Production Systems and Information Engineering*
Volume 13 (3), pp. 13–20.
https://doi.org/10.32968/psaie.2025.3.2

# Q-ÉRTÉK INICIALIZÁLÁS A HFRIQ-LEARNING RENDSZERBEN

TOMPA TAMÁS
Miskolci Egyetem
Informatikai Intézet
Általános Informatikai Intézeti Tanszék
tamas.tompa1@uni-miskolc.hu

KOVÁCS SZILVESZTER
Miskolci Egyetem
Informatikai Intézet
Általános Informatikai Intézeti Tanszék
szilveszter.kovacs@uni-miskolc.hu

**Absztrakt.** A Heurisztikusan gyorsított fuzzy szabályinterpoláció alapú Q-tanulás (HFRIQ-learning) módszer célja a szakértő által megadott tudásbázis injektálása a tanulási folyamatba (a tanulási fázis előtt), majd ezen kezdeti tudásbázis hangolása, pontosítása a tanulási folyamat során. A szakértő által megadható szabályok „ha-akkor” formátumúak, ahol a „ha” rész az állapot, az „akkor” rész pedig az ebben az állapotban preferált akció. A szakértői szabályrendszer HFRIQ-learning-rendszerbe történő injektáláshoz minden egyes szakértői szabályt „állapot-akció-Q-érték” formátumúra szükséges konvertálni, amely következtében szükséges a szabályok „Q-érték” részének meghatározása. A cikk célja egy kezdeti Q-érték-inicializálási módszer bemutatása, amely a szakértő által definiált minden egyes szabályra kezdeti Q-értéket határoz meg, amely következtében a szakértői szabályrendszer „állapot-akció-Q-érték” formátumban injektálható a tanulási folyamatba.

***Kulcsszavak:*** *megerősítéses tanulás, Q-learning, Fuzzy Q-learning, Q-érték-inicializálás, szakértői tudásbázis*

## 1. Bevezetés

A megerősítéses tanulás (Reinforcement Learning, RL) [5] egy olyan gépi tanulási paradigma, amelyben az ágens a környezetével való interakciók és az interakciók során kapott visszajelzések (megerősítések) alapján térképezi fel az adott probléma megoldását. Ezen módszereket trial-and-error (próbálkozás) típusú módszereknek is nevezik, mert akciók (cselekvések) próbálgatása során kapott megerősítési értékek alapján olyan jövőbeli akciók választására törekednek a módszerek, melyek maximalizálják a szerezhető jutalmakat.

Ilyen megerősítéses tanulási algoritmus a Q-tanulás (Q-learning) [16], illetve a SARSA [4] is, amelyek diszkrét állapot- és akciótér-felbontással rendelkeznek (azaz véges számú és diszkrét értékű állapot-akció érték lehetséges). Ezen algoritmusok közös jellemzője, hogy a Bellman-egyenlet [1] fixpontmegoldásait keresik számos iteráción keresztül és az adott állapotokban választható akciók hasznosságát egy Q-értékkel (hasznosságértékkel) jellemzik. A Q-érték az adott állapotban az adott akció végrehajtása melletti jóság értékeket határozza meg, tehát azt, hogy az adott akció végrehajtása mennyire eredményes („mennyit ér") az adott állapotban, figyelembe véve a jövőbeli várható jutalmakat is. Minden egyes lehetséges állapot-akció párra ez egy $Q(s_t, a_t)$ Q-függvényt eredményez, amely értékei általában (diszkrét állapot-akció tér esetében) egy Q-táblában tárolódnak, majd a tanulási fázis során pedig iteratív módon kerülnek frissítésre. A Q-értékek tehát ezen algoritmusok alapvető elemei, a tanulási folyamat során létrejött tudásbázist reprezentálják.

A Heurisztikusan gyorsított fuzzy szabály-interpoláció alapú Q-tanulás (HFRIQ-learning) [8] módszer esetében a rendszer tudásbázisát egy ritka (interpolált) fuzzy szabálybázis írja le, ahol a fuzzy szabályok konzekvens része a Q-érték, antecedense pedig az állapot-akció. Az állapot-akció formátumú szakértői szabályok tanulási folyamatba történő beillesztéséhez szükség van azok Q-értékének (nem szakértő által történő) meghatározására.

A cikk célja egy olyan Q-érték inicializálási módszer bemutatása, amely alkalmas a szakértő által definiált állapot-akció formátumú fuzzy szabályokra Q-értéket meghatározni, amely következtében azok már adaptálhatók a HFRIQ-learning-rendszer tanulási folyamatába.

## 2. A HFRIQ-learning megerősítéses tanulási módszer

A heurisztikusan gyorsított FRIQ-learning (Heuristically Accelerated Fuzzy Rule-Interpolation based Q-learning – HFRIQ-learning) [8] a FRIQ-learning (Fuzzy Rule-Interpolation based Q-learning) [13] [15] továbbfejlesztett változata, amely lehetőséget biztosít külső szakértői tudásbázis integrálására [6], valamint annak finomhangolására [10]. A módszer a Q-függvény reprezentálásához a „FIVE" (Fuzzy Rule Interpolation based on Vague Environment) [2] [3] szabályinterpolációs algoritmust alkalmazza, amely következtében módszer állapot-akció tere folytonos.

A rendszer tudásbázisa egyetlen $r_i$ ($i \in [1, m]$) szabályának formátuma az $m$ méretű $R$ szabálybázisban a következő [9][13]:

$$r_i\colon \boldsymbol{If}\ s_1\ is\ S_1^i\ \boldsymbol{And}\ s_2\ is\ S_2^i\ \boldsymbol{And}\ \ldots\ \boldsymbol{And}\ s_n\ is\ S_n^i\ \ \boldsymbol{And}\ a\ is\ A^i\ \boldsymbol{Then}\ \tilde{Q}(\boldsymbol{s}, a) = q^i \qquad (1)$$

ahol $S_j^i$ az $i$-edik ($i \in [1, m]$) szabály $j$-edik ($j \in [1, n]$) állapotdimenziójának fuzzy halmaza az $n$-dimenziós $\boldsymbol{S}$ állapottérben, $\boldsymbol{s} \in \boldsymbol{S}$ az $n$-dimenziós állapotmegfigyelés, $s_j$ a $j$-edik dimenziója az $\boldsymbol{s}$ állapotmegfigyelésnek, $A^i$ az $i$-edik szabály egydimenziós akcióuniverzumának $(U)$ fuzzy halmaza, $a \in U$ az akció,

$\tilde{Q}(\boldsymbol{s}, a)$ a FIVE FRI [2][3] által becsült Q-függvény, $q^i$ pedig az $i$-edik szabály konzekvense (Q-értéke).
A szakértői tudásbázis ($R_{expert}$) formátuma hasonló az (1) formula által meghatározott szabályokéhoz, azzal az eltéréssel, hogy a szakértői szabályok ($\hat{r}$) antecedense az állapot, konzekvense pedig az ebben az állapotban preferált akció [6]:

$$\hat{r}_i\text{: } \textbf{If } s_1 \boldsymbol{is}\ \hat{S}_1^i \ \boldsymbol{And}\ s_2 \boldsymbol{is}\ \hat{S}_2^i \ \boldsymbol{And}\ \dots \boldsymbol{And}\ s_n \boldsymbol{is}\ \hat{S}_n^i \ \boldsymbol{Then} \text{ a} = \hat{A}^i \tag{2}$$

ahol $\hat{r}_i$ az $i$-edik ($i \in [1, \hat{m}]$) szakértői szabály az $R_{expert}$ szabálybázisban, $\hat{S}_n^i = \left[\hat{S}_1^i, \hat{S}_2^i, \dots \hat{S}_n^i\right]$ az $i$-edik szakértői szabály $n$-dimenziós állapotmegfigyelése, $\hat{A}^i$ az ehhez az $\hat{S}_n^i$ állapotmegfigyeléshez tartozó akció, $i$ ($i \in [1, \hat{m}]$) pedig a szabály indexe az $\hat{m}$ méretű szakértői szabályrendszerben.

Annak következtében, hogy a szakértői szabályrendszer injektálható legyen a rendszerbe formátumának átalakítása szükséges „állapot-akció-Q-érték" formátumra. Ezután az átalakított szakértői szabályok antecedense az állapot-akció, konzekvense pedig egy becsült $\tilde{Q}_{init}$ érték lesz. A cikk célja ezen $\tilde{Q}_{init}$ érték számítási módjának bemutatása, amely a későbbi 3. fejezetben kerül részletesen bemutatásra.

A HFRIQ-learning tanulási folyamata egy kezdeti fuzzy szabálybázissal indul, amely a szakértői szabályok és az állapottér sarokpontjaiban elhelyezkedő, 0 konzekvens-értékű szabályok kombinációjával jön létre. A szakértői szabályok és a sarokponti szabályok között lévő esetleges ellentmondásokat a rendszer az ellentmondó szabályok összevonásával oldja fel. A tanulás során új szabályok akkor kerülnek beillesztésre a szabálybázisba, ha a Q-érték változása meghalad egy előre meghatározott küszöbértéket, és nincs az aktuális megfigyeléshez közeli, már létező szabály. Ha a Q-érték változása (frissítése) kicsi, akkor a meglévő szabályok konzekvensei frissülnek a következő összefüggés szerint [13] [15]:

$$\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \tilde{Q}^k(\boldsymbol{s}, a) + \Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) \tag{3}$$

$$\Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \alpha * \left(g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a') - \tilde{Q}^k(\boldsymbol{s}, a)\right) \tag{4}$$

ahol $\gamma \in [0,1]$ a leszámítolási tényező, $\alpha \in [0,1]$ a tanulási ráta, $q_i^{k+1}$ az $i$-edik szabály konklúziója a $(k+1)$-edik iterációban, $a$ pedig az $\boldsymbol{s}$-ben végrehajtott akció.
Az új megfigyelt állapot $\boldsymbol{s}'$, $g(\boldsymbol{s}, a, \boldsymbol{s}')$ a megfigyelt jutalom az $\boldsymbol{s} \to \boldsymbol{s}'$ állapotátmenetre, $\tilde{Q}^k$ és $\tilde{Q}^{k+1}$ pedig a $k$-adik és a $(k+1)$-edik iteráció FIVE FRI módszer által becsült Q-értéke [13][15]:

$$\tilde{Q}(\boldsymbol{s},a)=\begin{cases} q^i & \text{ha } (\boldsymbol{s},a)=(\boldsymbol{s}^i,a^i) \text{ valamennyi } i-re, \\ \sum_{i=1}^{m}\left(\left({q^i}/{(\delta_v^i)^\lambda}\right)\Big/\left(\sum_{j=1}^{m}{1}/{(\delta_v^j)^\lambda}\right)\right) & \text{egyébként} \end{cases} \tag{5}$$

ahol $q^i$ az $i$-edik ($i \in [1,m]$) szabály konklúziója, ($\mathbf{s}$, a) a megfigyelés, $\delta_v^i$ a skálázott távolság [3] az ($\mathbf{s}$, a) megfigyelés és az $i$-edik szabály ($\boldsymbol{s^i}, a^i$) antecedense között, $\lambda$ a Shepard-paraméter, $m$ pedig a szabályok száma.

Ha a Q-érték változása a frissítés során kicsi és van már létező szabály a megfigyelés közelében [8], akkor egy gradiens módszer alapú hangolási eljárás a megfigyeléshez legközelebb lévő szabálypont antecedensét és konzekvensét fogja hangolni [10] a következő összefüggések szerint [8] [10]:

$$\boldsymbol{s}_{k+1}=\boldsymbol{s}_k-\left(2*TDerror*\frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial\boldsymbol{s}}\right)*\alpha \tag{6}$$

$$a_{k+1}=a_k-\left(2*TDerror*\frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial a}\right)*\alpha \tag{7}$$

$$q_{k+1}=q_k-\left(2*TDerror*\frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial q}\right)*\alpha \tag{8}$$

ahol a $\boldsymbol{s}_{k+1}, a_{k+1}, q_{k+1}$ a gradiens módszer által meghatározott új állapot, akció- és Q-értékek, $\boldsymbol{s}_k, a_k, q_k$ a régi állapot, akció- és Q-értékek, $\alpha$ a gradiens-módszer tanulási rátája, $\frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial\boldsymbol{s}}, \frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial a}, \frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial q}$ a Q-függvény állapot, akció- és Q-érték szerinti parciális deriváltjai, a $TDerror$ értéke pedig a következő [8] [10]:

$$TDerror=g(\boldsymbol{s},a,\boldsymbol{s}')+\gamma*\max_{a'\epsilon U}\tilde{Q}^k(\boldsymbol{s}',a')-\tilde{Q}^k(\boldsymbol{s},a) \tag{9}$$

A tanulási folyamat során alkalmazott szabályhangolás következtében előfordulhat, hogy egyes szabálypontok közel kerülnek egymáshoz. Ilyen esetekben, mivel ezek a szabálypontok nagyon hasonló információt írnak le, a rendszer ezeket egyetlen szabállyá vonja össze a tanulási folyamat során, csökkentve ezáltal a szabálybázis méretét [8] [10]. További, a tanulási fázis után opcionálisan alkalmazható szabálybázis-csökkentő eljárásokat a [7] [11] [12] és [14] források mutatnak be.

A rendszer tanulási fázisa akkor tekinthető befejezettnek, ha már nem jön létre új szabály, a Q-értékek változása és a szabálypontok vándorlása minimális, illetve nem történik már újabb szabályösszevonás (nincsennek egymáshoz közeli szabályok).

## 3. Q-érték inicializációs módszer a szakértői szabályokra

Az előzetes tudásbázis a szakértő által szabályrendszer formájában kerül leírásra a (2) formula által meghatározott módon. A szabályrendszerben az antecedens a többdimenziós állapotuniverzumot, a konzekvens pedig az akciódimenziót jelöli, azaz ezen szabályok definiálásukkor nem rendelkeznek Q-értékkel. A HFRIQ-learning-módszer szabályrendszere a (1) formula szerint állapot-akció-Q-érték formátumú, ahol az antecedens a többdimenziós állapot és az akció, a konzekvens pedig a Q-érték.

Annak érdekében, hogy a szakértői szabályok a HFRIQ-learning szabályrendszerébe illeszthetőek legyenek, meg kell határozni a szakértői szabályok kezdeti Q-érték konzekvensét. A szakértői szabályok akciókonzekvensei az antecedens oldalra kerülnek, majd az új konzekvensük pedig ezen Q-érték lesz. Ezt a folyamatot még a tanulási fázis megkezdése előtt kell megvalósítani. A kezdeti Q-érték meghatározási módszer célja tehát még a tanulási fázis előtt, a szakértői szabályrendszer minden egyes szabályára becsült Q-érték $(\tilde{Q}_{init})$ inicializálása, azaz a kezdeti Q-függvény meghatározása. Ez a szakértői heurisztikából létrehozott kezdeti Q-függvény lesz hangolva a tanulási folyamat során.

Feltételezve, hogy a szakértő által megadott szabályok megkérdőjelezhetetlenül helyesek, az azokra meghatározott Q-értékeknek relatívan magasnak (és lehetőleg 0-tól eltérőnek) kell lenniük. A relatívan magas érték egy kezdeti becslés, amely a környezet által maximálisan adható jutalom $(g_{max})$ ismeretében a következők alapján határozható meg:

Jelöljük a szakértői szabályok kezdeti becsült Q-értékét $\tilde{Q}_{init}$-tel. Ez a kezdeti Q-érték-becslés pedig legyen a maximálisan szerezhető Q-érték $(\tilde{Q}_{max})$ érték $\eta$-ed ($\eta \in [0,1]$) része:

$$\tilde{Q}_{init} = \eta * \tilde{Q}_{max} \tag{10}$$

Ahol $\eta$ azt határozza meg, hogy a maximálisan szerezhető, becsült Q-érték mekkora része legyen a szakértői szabályok tényleges konzekvense értéke (Q-értéke). A $\tilde{Q}_{max}$ értékét a következőképpen lehet kifejteni a (*4*) frissítési formula alapján:

$$\tilde{Q}_{max} = \lim_{k\to\infty} \tilde{Q}^{k+1}(\boldsymbol{s}^*, a^*)$$
$$\tilde{Q}_{max} = \lim_{k\to\infty} (\tilde{Q}^{k}(\boldsymbol{s}^*, a^*) + \alpha * g(\boldsymbol{s}^*, a^*, \boldsymbol{s}^*) + \gamma * \tilde{Q}^{k}(\boldsymbol{s}^*, a^*) - \tilde{Q}^{k}(\boldsymbol{s}^*, a^*))) \tag{11}$$

Ahol legyen $\tilde{Q}^k(\boldsymbol{s}^*, a^*) = \max_{a' \in U} \tilde{Q}^k(\boldsymbol{s}^*, a')$ és $g(\boldsymbol{s}^*, a^*, \boldsymbol{s}^*) = \max_{\boldsymbol{s}\in\boldsymbol{S}, a\in U} g(\boldsymbol{s}, a, \boldsymbol{s}') = g_{max}$. $\alpha$ a tanulási ráta, $\gamma$ a diszkontálási tényező, $g(\boldsymbol{s}, a, \boldsymbol{s}')$ a jutalom értéke az $\boldsymbol{s} \to \boldsymbol{s}'$ állapotátmenetre, $\tilde{Q}^k$ a $k$-adik, $\tilde{Q}^{k+1}$ pedig a $k+1$-edik iteráció becsült konklúziója. A $g_{max}$ paraméter, amely szintén a szakértő által meghatározott és a környezet által adható maximális megerősítés értékét jelöli. Ez a (*11*) összefüggés azt mutatja meg,

hogy miközben az iterációk száma ($k$) közelít a végtelenhez a Q-érték hova konvergál.
A (*11*) összefüggést tovább fejtve a következő formulához jutunk:

$$\begin{aligned}\tilde{Q}_{max} &= \lim_{k\to\infty}(\tilde{Q}^k(\boldsymbol{s}^*, a^*) + \alpha * (g_{max} + (\gamma - 1) * \tilde{Q}^k(\boldsymbol{s}^*, a^*))) = \\ &= \tilde{Q}^k(\boldsymbol{s}, a) + \alpha * g(\boldsymbol{s}, a, \boldsymbol{s}') + \alpha * (\gamma - 1) * \tilde{Q}^k(\boldsymbol{s}', a') = \\ &= \frac{\alpha * g_{max}}{-\alpha * (\gamma - 1)} = \frac{g_{max}}{1 - \gamma}\end{aligned} \tag{12}$$

A fenti levezetés alapján a szakértői szabályok kezdeti becsült Q-értéke $(\tilde{Q}_{init})$ a környezet által adható maximális megerősítés ($g_{max}$) és a 1-diszkontálási faktor ($\gamma$) hányadosa:

$$\tilde{Q}_{init} = \eta * \frac{g_{max}}{1-\gamma}\text{, ahol } \gamma < 1 \tag{13}$$

Ahol $\tilde{Q}_{init}$ a becsült, maximálisan elérhető Q-érték a környezet által maximálisan adható jutalom $g_{max}$ ismeretében, $\eta \in [0,1]$ a $\tilde{Q}_{init}$ skála faktora, amely azt határozza meg, hogy a számított $\tilde{Q}_{init}$ érték mekkora része (százaléka) kerüljön figyelembevételre, $\gamma$ pedig a diszkontálási tényező. Az így kiszámított $\tilde{Q}_{init}$ érték a szakértői szabályrendszer minden egyes szabályára, azaz annak konzekvens részére egyforma értékű.

A szabályrendszer formája az előzetes Q-érték meghatározási módszer alkalmazása után az (1) összefüggés mintájára a következőképpen módosul:

$$\hat{r}_i\text{: } \textbf{If } s_1 \textbf{ is } \hat{S}_1^i \textbf{ And } s_2 \textbf{ is } \hat{S}_2^i \textbf{ And } \ldots \textbf{ And } s_n \textbf{ is } \hat{S}_n^i \textbf{ And a} = \hat{A}^i \textbf{ Then } \tilde{Q}(\boldsymbol{s}, a) = \tilde{Q}_{init} \tag{14}$$

## 4. Összefoglalás

A cikk egy olyan Q-érték inicializálási módszert mutatott be, amely lehetővé teszi a HFRIQ-learning-rendszerben megadott szakértői szabályokra történő kezdeti Q-értékek meghatározását. Az eljárás célja, hogy az állapot-akció formátumú szakértői szabályokat a módszer hatékonyan integrálja a tanulási folyamatba. Ennek során a szakértői szabályok állapot-akció-Q-érték formára kerülnek átalakításra, majd a Q-érték rész kezdeti értékének meghatározása a javasolt módszer által történik, amely a szakértő által megadott maximális megerősítési értéken alapul. A bemutatott módszer hozzájárul a szakértői tudásbázis HFRIQ-rendszer tanulási folyamatába történő injektálásához és a szakértői tudás hatékony felhasználásához.

## Irodalomjegyzék


[1] Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

[2] Kovács, Sz., Kóczy, L. T. (1997). The use of the concept of vague environment in approximate fuzzy reasoning. *Fuzzy Set Theory and Applications*. Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Bratislava, Slovak Republic, vol. 12, 169–181.

[3] Kovács, Szilveszter (2006). Extending the fuzzy rule interpolation "five" by fuzzy observation. *Computational Intelligence, Theory and Applications: International Conference 9th Fuzzy Days in Dortmund, Germany, Sept. 18–20, 2006 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-34783-6_48

[4] Rummery, G. A., Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. CUED/F-INFENG/TR 166, Cambridge University, UK.

[5] Sutton, Richard S., and Andrew G. Barto (2018). *Reinforcement learning: An introduction*. MIT press. https://doi.org/10.1017/s0263574799211174

[6] Tompa, Tamás, and Szilveszter Kovács (2020). Applying Expert Heuristic as an a Priori Knowledge for FRIQ-Learning. *Acta Polytechnica Hungarica*, 17, 4. https://doi.org/10.12700/aph.17.4.2020.4.2

[7] Tompa, Tamás, and Szilveszter Kovács (2017). Clustering-based fuzzy knowledgebase reduction in the FRIQ-learning. *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, IEEE. https://doi.org/10.1109/sami.2017.7880302

[8] Tompa, Tamás, and Szilveszter Kovács (2022). Heuristically accelerated FRIQ-learning. *20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2022)*, IEEE. https://doi.org/10.1109/SISY56759.2022.10036311

[9] Tompa, Tamás, and Szilveszter Kovács (2024). Integrating Expert Knowledge into Fuzzy Reinforcement Learning. *2024 IEEE 18th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, IEEE. https://doi.org/10.1109/saci60582.2024.10619888

[10] Tompa, Tamás, and Szilveszter Kovács (2024). Knowledge Base Optimization of the HFRIQ-Learning. *Acta Polytechnica Hungarica*, 21, 10. https://doi.org/10.12700/aph.21.10.2024.10.6

[11] Tompa, Tamás, and Szilveszter Kovács (2023). Tudásbázis redukálás a heurisztikusan gyorsított FRIQ-learning rendszerben. *Production Systems and Information Engineering*, 11, 2, 1–12. https://doi.org/10.32968/psaie.2023.2.1

[12] Vincze, Dávid, Alex Tóth, and Mihoko Niitsuma (2020). Antecedent redundancy exploitation in fuzzy rule interpolation-based reinforcement learning. *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. https://doi.org/10.1109/aim43001.2020.9158875

[13] Vincze, Dávid, and Szilveszter Kovács (2009). Fuzzy rule interpolation-based Q-learning. *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, IEEE. https://doi.org/10.1109/saci.2009.5136311

[14] Vincze, Dávid, and Szilveszter Kovács (2015). Rule-base reduction in Fuzzy Rule Interpolation-based Q-learning. *Recent Innovations in Mechatronics*, 2, 1–2, 1–6. https://doi.org/10.17667/riim.2015.1-2/10.

[15] Vincze, Dávid (2013). *Fuzzy Rule Interpolation-based Q-learning*. PhD dissertation.

[16] Watkins, C. J. C. H., Dayan, P. (1992). Q-learning. *Machine Learning*, vol. 8 (3/4), 279–292.