https://doi.org/10.63684/dk.2026.09.03

JUHÁSZ LEVENTE ZSOLT *

Megerősítéses tanulás a pszichológiában és a mesterséges intelligenciában

Összefoglalás: A megerősítéses tanulás (reinforcement learning, RL) tudományos szempontból azt vizsgálja, hogyan képesek az állatok és a gépek viselkedésüket úgy alakítani, hogy eközben a lehető legtöbb jutalmat érjék el. Az RL kutatás gyökerei egészen a pszichológiában végzett, az instrumentális tanulásra vonatkozó korai munkákig nyúlnak vissza.
A modern RL azonban egy erősen interdiszciplináris terület, amely a számítástechnika, a gépi tanulás, a pszichológia és az idegtudomány közös metszéspontján helyezkedik el. Az előadás összefoglalja a terület legfontosabb matematikai alapelveit, mint például a felfedezés/kiaknázás dilemmáját (exploration/exploitation dilemma), az időbeli különbségen alapuló tanulást (TD-tanulás) (temporal-difference learning), a Q-tanulást (Q-learning), valamint a modellalapú és modellmentes tanulás közötti különbségeket. Végül áttekinti a pszichológia és az idegtudomány területén felmerülő számos nyitott kérdést.
Kulcsszavak: Megerősítés tanulás, kondicionálás, mesterséges intelligencia.

Abstract: Reinforcement learning (RL) is the scientific study of how animals and machines can shape their behavior to maximize rewards. The roots of RL research go back to early work in psychology on instrumental learning.
Modern RL, however, is a highly interdisciplinary field that lies at the intersection of computer science, machine learning, psychology, and neuroscience. The presentation will summarize the most important mathematical principles of the field, such as the exploration/exploitation dilemma, temporal-difference learning (TD-learning), Q-learning, and the differences between model-based and model-free learning. Finally, it will review a number of open questions both in psychology and neuroscience.
Keywords: Reinforcement learning, conditioning, artificial intelligence.

* *Dunaújvárosi Egyetem, Tanárképző Központ, egyetemi docens*
Email: juhaszle@uniduna.hu
ORCID: 0009-0006-0240-2937

A tanulás és alaptípusai

Tankönyvi definíció szerint [1] a tanulás a viselkedésnek a tapasztalatok következtében kialakuló viszonylag tartós megváltozása. Hagyományosan két alapformát különböztetünk meg:
– asszociációs tanulás,
– nem asszociációs tanulás.

A nem asszociációs tanulás jellemzője, hogy csak egy bizonyos ingerre vonatkozik, általában rövidebb ideig tart. Két alapformája van, amelyek univerzálisnak tekinthetők az állatvilágban. Az egysejtű szerveződések, sőt egyes növényfajok esetén is megfigyelhetők. Az egyik alapforma a habituáció („megszokás"), amelyet akkor tapasztalhatunk, ha valamely ingerre adott válasz gyakorisága csökken. A másik forma a szenzitizáció („érzékenyülés"), ekkor valamely ingerre a tapasztalatok alapján egyre erősebb válasz érkezik. Ez főleg ártalmas ingerek esetén figyelhető meg [2].

Az asszociációs tanulás már többsejtű állatok sajátja, működéséhez már bizonyos komplexitást meghaladó idegrendszer szükséges. Események, ingerek közötti kapcsolatok elsajátítása, vagyis asszociációkat alakítunk ki. A pszichológiában az asszociációs tanulás fogalma alatt a kondicionálást, és a közvetlen kapcsolatképzésen túlmutató komplex tanulást értik. A kondicionálás két formája a klasszikus kondicionálás és az instrumentális (operáns) kondicionálás. A klasszikus kondicionálás során események közötti kapcsolatok alakulnak ki. Az instrumentális kondicionálás során az egyedek azt sajátítják el, hogy bizonyos viselkedéses válaszok bizonyos következményekkel járnak. A komplex tanulás pedig magasabb rendű, az asszociációelvvel teljesen nem magyarázható tanulás (mentálist térkép, kategorizáció, mintázatészlelés stb.)

A tanulás megközelítései [1]:
– *Behaviorista:* az ingerek, illetve ingerek és válaszok közötti kapcsolatok elsajátításának vizsgálata. Az egykor mainstream viselkedéslélektani felfogás szerint minden tanulás alapköve a kondicionálás, és a minden tanulásra ugyanazok a törvények vonatkoznak tartalomra és tanulótól függetlenül (Skinner). Az irányzat követői főleg állatok tanulását vizsgálták. A komplex tanulás jelenségét zárójelbe tették, nem magyarázzák.

[1] Smith, E. E.–Nolen-Hoeksema, S.–Fredrickson, B. L.–Loftus, G. R. (2005): Atkinson–Hilgard: *Pszichológia.* Budapest: Osiris.

[2] Gluck, M. A.–Mercado, E.–Myers, C. E. (2019): *Learning and Memory: From Brain to Behavior*. New York: Worth Publishers.

– *Kognitív:* a tanuló egyedek előzetes ismereteit is figyelembe veszi. Az alkalmazott stratégiákat és szabályok is érdekesek.
– *Biológiai:* a tanulás neurológiai háttere, fajra jellemző tanulási formák vizsgálata (etológia és élettan).
– *Gépi tanulás*: a mesterséges intelligencia részterülete, célja tanulni képes mesterséges információs rendszerek létrehozása.

Klasszikus kondicionálás [3]

Lényege egy korábban semleges inger egy másik ingerrel való ismételt együttjárása következtében asszociálódik egy másik ingerhez. A jelenség feltárása és leírása az emésztés élettanát kutató Pavlov megfigyelései alapozták meg. A kutyakísérletekben, az állatok nyálelválasztása az etetőtál látványára fokozódik. Később más ingerekhez (pl. vizsgálat kezdetét jelző csengőhang) is tudták asszociáltatni a nyáltermelést. A kutatások során kialakult terminológia szerint a Feltétlen inger (UCS) egy adott választ automatikusan, tanulás nélkül, általában reflexesen kiváltó inger (pl. étel látványa, íze). A Feltétlen válasz (UCR) a feltétlen ingerre adott eredeti válasz, amely egy eredetileg semleges ingerhez kapcsolva feltételes válasszá alakítható át. (Pl. nyáladzás) A Feltételes inger (CS) egy korábban semleges inger, amely egy feltétlen ingerhez asszociálva feltételes választ eredményez. (pl. a hanginger, vagy fény). A Feltételes válasz (CR): A korábban semleges feltételes ingerhez kapcsolódó tanult vagy szerzett válasz (eredetileg feltétlen válasz volt: nyáltermelés).

Instrumentális kondicionálás [2]

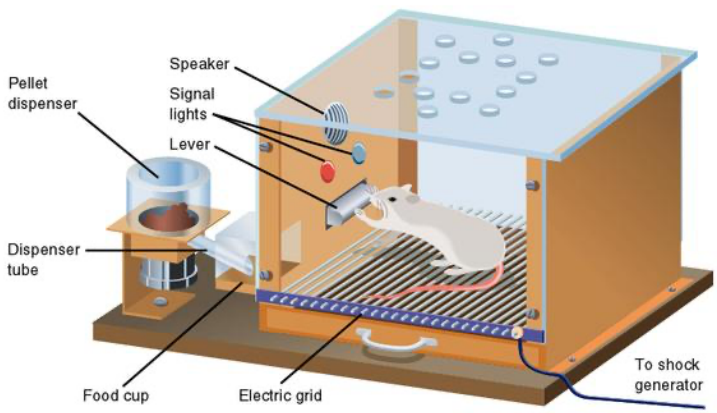
Az élőlények környezetükre igyekeznek hatni (instrumentális), azt befolyásolni (emiatt operáns kondicionálásnak is nevezik). Megtanulják, hogy viselkedésüknek következményei vannak. Általában új válaszformák elsajátítására alkalmazzák pl. állatok idomítása esetén. Valamely viselkedésforma újbóli megjelenése attól függ milyen hatása volt, közelebb vitt-e a célhoz. Ez az effektus törvénye (Thorndike, 1898), ami darwinista tanuláselmélet alapja. Thorndike úttörő megfigyelésében a ketrecbe zárt macskák először próba-szerencse (try and error) viselkedéssel nyitják ki a zárat.

[2] Gluck, M. A.–Mercado, E.–Myers, C. E. (2019): *Learning and Memory: From Brain to Behavior*. New York: Worth Publishers.

[3] Domjan, M.–Delamater, A. R. (2023): *The Essentials of Conditioning and Learning.* New York: American Psychological Association.

A próbák során egyre rövidebb ideig tart az irreleváns viselkedés fázisa; a végén szinte azonnal kinyitja, rálép a reteszre. Az effektus törvénye szerint a tanulási folyamat során a véletlenszerű viselkedéselemek közül azok választódnak ki, melyek pozitív következményekkel (jutalom) járnak. Az operáns tanulás összefüggéseinek megismeréséhez nagymértékben hozzájárultak B. F. Skinner szisztematikus és módszertanilag megalapozott állatkísérletei. Eszköze az ún. Skinner-doboz volt.

1. ábra. Skinner-doboz



Forrás: https://forgos.uni-eszterhazy.hu/wp-content/tananyagok/fs_komm_egyetemi/obj/ie_0039_0_0_0/0039_0_0_0_forras.htm)

A dobozban egy pedál az etetőtál felett helyezkedik el, plusz egy lámpa, amit a kísérletvezető vezérelt. A pedálnyomás alapszintje: a „naiv" állat pedálnyomásának kezdeti gyakorisága (amikor még nincs megerősítés). A tanítási szakaszban, ha az állat a pedálra lépett ételt kapott. A pedálnyomás gyakorisága ennek hatására általában jelentősen megnőtt. A válasz gyakoriság változása jelezte a tanulás sikerességét. A kioltás jelensége, ha folyamatosan elmarad a megerősítés, ekkor a válaszgyakoriság újra lecsökken. A válasz gyakoriságot befolyásoló ingereket hatásuk alapján megerősítő ingereknek vagy büntetésnek nevezhetjük. Megerősítés az a folyamat, amely során egy appetitív inger megjelenése (pozitív megerősítés) vagy az averzív inger megszűnése (negatív megerősítés) növeli a viselkedés valószínűségét A büntetés során egy averzív inger megjelenése (pozitív büntetés) vagy az appetitív inger megszűnte (negatív büntetés) csökkenti a viselkedés valószínűségét.

A kondicionált válasz generalizálódhat és/vagy diszkriminálódhat. Generalizáció esetén valamilyen ingerre adott választ korábban már megerősítették. Ugyanezt a választ hasonló ingerekre is mutatja az egyed. Diszkrimináció ezzel ellentétes folyamat. Ekkor csak a kívánt ingerekre adott választ erősítik meg. Annál hatékonyabb a diszkriminációs inger, minél inkább különbözik a gátlandó ingertől, amilyen mértékben megléte az ingerre adott válasz megerősítését előrejelzi. Instrumentális kondicionálás esetén különböző megerősítési terveket lehet alkalmazni: a kívánt viselkedések milyen arányban és milyen frekvenciával (idői gyakorisággal) kapjon megerősítést – idői és aránytervek. A részleges megerősítés (nem minden válasz után) általában sokkal hatékonyabb, sokkal nehezebb kioltani

Pl. a galamb, ha csak óránként 5 megerősítést kap, akkor is 6000-szer koppint az élelemért [1]. Skinner (behaviorista) szerint a kontiguitás (együttjárás), az inger-válasz időbeli érintkezése a fontos. A válasz akkor kondicionálódik, ha a megerősítés közvetlenül követi. Seligman (kognitív) felfogása szerint a válasz csak akkor kondicionálódik, ha az egyed úgy érzékeli, hogy a megerősítés az ő válaszától függ (perception of control).

Gépi tanulás

Tulajdonképpen egy szoftver által megvalósított statisztikai tanulás. Tom Mitchell [4] klasszikus definíciója szerint: „Azt mondjuk, hogy a számítógép tanul egy T feladatban a P teljesítmény metrika szerint az E tapasztalatából, ha a teljesítménye a T feladatban, amit P-vel mérünk javul az E tapasztalattal."

A gépi tanulás fontosabb megközelítései [5]

Felügyelt tanulás. Ekkor a program tanítópéldát kap, ami a bemeneti inputot és az elvárt inputot is tartalmazza. A program az elvárt és becsült kimenet különbségén alapuló hibaértéket csökkenti optimalizációval. Ide sorolják a regressziós, kategorizációs/osztályozási feladatokat.

Tanulás felügyelet nélkül. Ennél a típusnál nincsenek elvárt output-input asszociációk és tanítópéldák, a program feladata az adatokban rejlő összefüggések feltárása. Tipikus példái a klaszterezés, és a dimenziócsökkentés (pl. PCA , faktoranalízis).

[1] Smith, E. E.–Nolen-Hoeksema, S.–Fredrickson, B. L.–Loftus, G. R. (2005): Atkinson –Hilgard: *Pszichológia.* Budapest: Osiris.

[4] Mitchell, T. (1997): *Machine Learning.* New York: McGraw-Hill.

[5] Géron, A. (2026): *Hands-On Machine Learning with Scikit-Learn and PyTorch.* New York: O'Reilly.

[6] Sutton, R. S.–Barto, A. G. (2018): *Reinforcement Learning.* New York: Bradford Books.

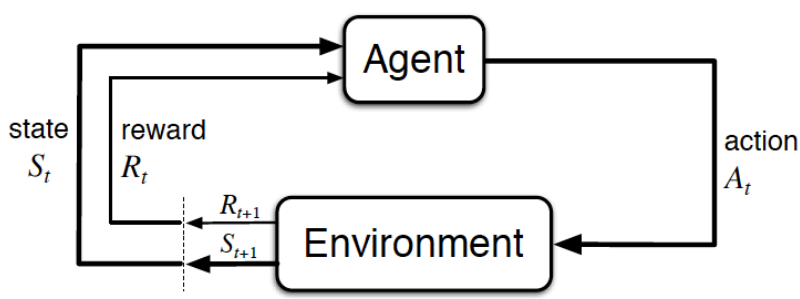
Az *önfelügyelt tanulás* átmenetet képez ez előző két típus között. A tanító halmazban nincsenek külön bemeneti és elvárt kimeneti ingerek, hanem a rendszer saját kimenete szolgál egy következő lépésben bemenetként. Általában szekvenciális, idősoros adatok feldolgozásának megtanulása lehetséges így. Típikus példája az ún. autoregressive training, amelyek a transformer achitektúrájú Nagy Nyelvi Modellek (LLMs) tanításának az eszköze.

Megerősítéses tanulás. Egy a környezetével akcióin keresztül interakcióban lévő szoftverrendszer tanulási formája. Célja az összesített megerősítések (hasznosságok) maximalizálása. Ezt, mivel a természetes tanulás operáns típusának a gépi megfelelője, egy külön bekezdésben, részletesebben tekintjük át.

Megerősítéses tanulás [6]

Ilyenkor a programnak nem egyetlen döntést, hanem döntések sorozatát kell meghoznia. Az elemi döntések helyességére nem feltétlen kap visszajelzést, tipikusan csak a cselekvéssorozat végén kap jutalmat/büntetést.

2. ábra. A megerősítéses tanulás absztrakt modellje



A megerősítéses tanulás modelljében egy absztrakt ágens a szoftver által szimulált környezetben hajt végre akciókat. Robotikai feladatok és önirányította autonóm fizikai rendszerek esetén a tanulási eredménye később valós fizikai környezetben is alkalmazható. Az absztrakt környezet a Markov-féle Döntési Folyamatnak nevezett ötössel írható le: (S, A, R, P, γ).

Ebben:
– *S az állapottér:* s_1, s_2, … Az ágens ezekbe a környezeti állapotokba kerülhet.
– *A akció tér:* a_1, a_2, …. Az egyes környezeti állapotokban az ágens által végrehajtható akciók halmaza.
– R a megerősítések halmaza: r_1, r_2, …Bizonyos állapotokban végrehajtott akciók valamilyen megerősítéssel járhatnak együtt. Az ágens ezeket összegzi.
– *P átmeneti mátrix.* Egylépéses dinamika:
$p(s', r|s, a) = P(S_{t+1} = r', R_{t+1}=r| S_t = s, A_t = a)$ minden s', r, s, a-ra.
Valamely állapotban végrehajtott akció eredménye, csak az aktuális állapottól függ.
– *γ, diszkont ráta:* a rendszer a jövőbeli állapotok lehetséges értékét a γ tényezővel diszkontáltan veszi figyelembe.

Sztochasztikus policy az a valószínűségi összefüggés, $\pi(s,a) = p(a|s)$, amely kifejezi, hogy *S* állapotban az ágens milyen valószínűséggel választja az *A* akciót. A feladatban az ágens célja egy olyan stratégia, az egyes állapotokhoz tartozó akciók („policy") megtanulása, amely a legnagyobb várható hasznosságot, az akciósorozat folyamán a legtöbb jutalmat vonja maga után. A gépnek nagyon sokszor kell próbálkoznia, hogy megtalálja a helyes lépéssorozatot. Adott policy esetén meghatározható, hogy milyen várható nyereséget képes az ágens felhalmozni, minden állapot esetén. Ezt az összefüggést, az adott policy-hez tartozó **állapot-érték függvény**nek nevezzük: v_π (s). Minden állapothoz a kumulatív megerősítések várható értékét rendeli. Egy adott állapotból az *A* akciót végrehajtva, majd továbbiakban a π policy-t követve az **akció-érték függvény**, q_π(s, a) állapítható meg: ami minden állapothoz a kumulatív megerősítések várható értékét rendeli adott policy követése esetén.

Optimális policy (π^*) esetén v_π és q_π minden állapotra (és akcióra) a legjobb várható értéket adja (az ehhez tartozó állapot- és akció-értékfüggvények: v^*, q^*). A rendszer ezt az optimális ütemezést keresi. Ennek egyik módja, hogy meg kell találni q^* függvényt (ez egyben v^* is megadja) amiből az optimális policy kiszámítható. De az újabb módszerek, melyek a mélytanuláson alapuló neurális hálózatokat alkalmaznak, már nem igénylik ezeknek a függvényeknek az explicit becslését.

A *Monte-Carlo (MC) módszerek* kezdetben teljesen véletlen policy-ból indulnak ki, és az akció-érték függvény táblázatot (a q^* függvény becslésére) frissítik az epizódok átlagai alapján. A módosított q^* alapján módosul az akció választás is (policy). A szimulációk során az ágensnek meg kell találni az Exploitation - exploration egyensúlyt, azaz a már ismert, járt utat (jelenlegi legjobb policy) vagy járatlant (az aktuálisan legjobb policy-től eltérő akció) válassza?

A *Temporal difference (TD, idői különbség)* módszerek a tanuló algoritmusok másik nagy, és hatékonyabb családját jelentik. Az ágens minden lépésben frissíti a q-táblázat értékeit. Nem kell az epizód végét megvárnia. A kapott megerősítés és a következő lehetséges lépés a q-táblázat értéke alapján meghatározható. Ezután történik a következő lehetséges lépés kiválasztása. A legismertebb TD-algoritmusok a Sarsa, Sarsa-max (Q-learning).

Deep Q-learning. Ekkor a Q-hasznosságot egy egyszerű táblázat helyett egy neuronháló fogja tanulni. A háló inputja az aktuális állapot, ez egyes kimenő neuronjai pedig a lehetséges akcióknak felelnek meg, azok hasznosságát becsülik meg.

*Value-based módszerek*nek nevezik az optimális q*-kereső eljárásokat. Gyakran konvolúciós hálózat alkalmaznak erre. Policy-based módszerek közvetlenül π*-t (optimális) akció stratégiát tanulják. Bizonyos szempontból egyszerűbbek, és folytonos akció térben is hatékonyak (pl. policy gradient módszer)

Alkalmazások

A megerősítéses tanulás legsikeresebb alkalmazásai a játékhoz kötődnek: Például sakk, Atari játék, Go, Starcraft. A Deep-Q learning módszer változataira alapuló eljárást alkalmaznak. A tanulás inputja az aktuális képernyőkép kimenete és a lehetséges joystick mozdulatok (Atari) hasznossága [7]. Egy mély konvolúciós neuronhálót alkalmaztak. Alpha Go (Google) már kettős mély hálózatot is tartalmazott [8]. Az ún. policy network megbecsülte, hogy mit érdemes lépni, míg a value network a megtett lépés hasznosságára adott becslést. Először felügyelt tanítást alkalmaztak, majd a rendszer önmagával játszott és ezeket értékelte ki megerősítéses tanulás módszerével. Az optimális stratégia kialakulásához fakereső algoritmust is használtak: Monte Carlo Tree Search (MCTS). A módszer hatékonyságát tükrözi, hogy az AlphaGo 2015-ben legyőzte az aktuális világbajnokot is. Továbbfejlesztett változata, az Alpha Go Zero [9] kizárólag önmaga ellen játszva tanult, már az emberi játékosok számára is meglepő stratégiákat volt képes alkalmazni.

Dopamin és megerősítéses tanulás

RL Temporal Difference, a Q-tanulás (idői különbségen alapuló tanulás) paradigma értelmezhető a dopaminnal kapcsolatos idegtudományi eredmények alapján. A dopamin az idegrendszer sejtei közötti kommunikáció fontos neurotranszmittere; dopamin alapú agyi kapcsolatok működési zavarát számos mentális probléma esetén kimutatták. A Reward Prediction Error (RPE) jelenti azt, hogy mennyire különbözik az elvárt jutalom és a valós jutalom. Ha a megerősítés pozitívabb volt, akkor nőtt a dopamin-aktivitás, és ez megerősíti az akciót; ha alacsonyabb, akkor csökken a DA-aktivitás, és gyengíti az akciót.

[7] Mnih, V. e. (2015): Human-level control through deep reinforcement. *Nature,* 518., pp. 529–533.

[8] Silver, D. E. (2016): Mastering the game of Go with deep neural networks and tree search. *Nature,* 529., pp. 484–489.

[9] Silver D, E. A. (2017): Mastering the game of Go without human knowledge. *Nature,* 550., pp. 354-359.

A dopamin felszabadulása az agyban, különösen a ventrális tegmentális terület (VTA) és a nucleus accumbens közötti pályákon, közvetlenül tükrözi a jutalom-előrejelzési hibát (RPE). fMRI eredmények, striatum, prefrontal cortexben támasztják elő ezt az elképzelést [10].

A komputációs pszichiátria egy interdiszciplináris terület, amely matematikai és számítógépes modelleket használ az agy funkciói és az információfeldolgozás megértésére, lehetővé téve a mentális zavarok mögött meghúzódó kognitív és biológiai mechanizmusok precíz azonosítását [11]. Célja, hogy kvantitatív, mérhető paraméterekkel írja le az egészséges és kóros agyműködés közötti különbségeket, ezzel javítva a diagnózist és a személyre szabott kezeléseket. A megerősítéses tanulás modellje segítségével számos pszichopatológiai tünet értelmezhető [12]:

1. táblázat

Mentális probléma	RL modellhez kapcsolódó hipotézisek	Tünet magyarázata
Depresszió és szorongás	Emelkedett Büntetés Tanulási Ráta (nagyobb érzékenység a negatív kimenetekre) és/ vagy Csökkent Jutalmazási Érzékenység (kisebb válasz a pozitív kimenetekre).	Megmagyarázza a negatív affektív torzítást, a hibákon való fokozott rágódást és az anhedóniát (képtelenség az öröm megtapasztalására).
Skizofrénia	Abnormális Jutalom-Előrejelzési Hiba jelzés, amely néha a predikciós hibán alapuló tanulásra való túlzott támaszkodáshoz vezet.	Alátámaszthatja a téveszmék kialakulását (abnormális vélekedés frissítés) és a motiváció/akaratgyengeség hiányát (negatív tünetek).
Függőség (Addikció)	Az Azonnali Jutalom Növekvő Értékelése és zavar a Modell-Független (szokásos) vs. Modell-Alapú (célvezérelt) kontroll egyensúlyában.	Megmagyarázza az ellenőrzött drogfogyasztásról a negatív hosszú távú következmények ellenére is fennálló kényszeres, szokásalapú viselkedésre való áttérést.

[10] Masset, P.–Geshman, S. J. (2025): Reinforcement learning with dopamine: A convergence of natural and artificial intelligence. In: S. J. Cragg, –M. E. Walton: *The Handbook of Dopamine.* pp. 305–318. New York: Academic Press.

[11] Seriés, P. (2020): *Computational Psychiatry.* Boston: The MIT Press.

[12] Heinz, A. (2017): *A New Understanding of Mental Disorders.* Boston: The MIT Press.

Impulzivitás/ADHD	Meredekebb idői diszkontálás (a jövőbeni jutalmak alulértékelése) vagy fokozott döntési zaj (gyenge választási konzisztencia).	

A megerősítés tanulás további legfontosabb alkalmazásai [13]:
– Játékos szimuláció, játékok tesztelése,
– Robotika, önirányító /autonóm rendszerek,
– Pénzügyek, algoritmusos kereskedés, portfólió management, csalás detektálás,
– Ajánló rendszerek (recommender systems),
– Smartenergia és forrás management, smartgrid-szabályozás,
– Egészségügy, személyes orvostudomány; dinamikus kezelési rezsimek, forrásallokáció,
– Nagy Nyelvi modellek (LLM) tréningje, finomhangolása (fine tuning).

Gépi versus emberi tanulás

A gépi tanulás egyelőre nagyon lassú és nagyon nagy mintára (tanuló halmaz) van szükség. A gépi program nem vagy nehezen generalizál; 'metalearning' még nem jellemző rá. Egyelőre Általános Mesterséges Intelligenciáról nem beszélhetünk, csak szűk tudás/képességterületen hatékony, de ott rendszerint legyőzi az embert, meghaladja képességeinket [14; 15]. Az emberi tanuláshoz képest a gépi tanulás sokkal energiaigényesebb, kb. 2000-szer hatékonyabb a mi agyunk [16]. Ezeknek a problémáknak a megoldása a Mesterséges Intelligencia kutatás jövőbeli feladatait képezik.

[13] Murphy, K. P. (2026): *Reinforcement Learning: An Overview.*

[14] Norvig, P.–Russell, S. (2023): *Mesterséges intelligencia 1–2*. Budapest: Taramix.

[15] Chollet, F. (2026): *Deep Learning with Python.* New York: Manning.

[16] LeCun, Y. (2025): *A mesterséges intelligencia és a mélytanulás forradalma.* Budapest: TypoTex.