FILOLÓGIA.HU • XVI. évfolyam (2025) 3–4. szám 134–146. • DOI: 10.59648/filologia.2025.3-4.10

Kolláth Mihály Gábor

Anyanyelv, MI és jog

Hogyan tud majd magyarul a mesterséges intelligencia?

Language, AI and Law

How Much Hungarian Will Artificial Intelligence Speak?

A mesterséges intelligencia korunk egyik legtöbb lehetőséggel kecsegtető, ugyanakkor számos kockázattal fenyegető technológiája. Az utóbbi években életünk irányításának, szervezésének mind nagyobb részét engedtük át ezen technológiának. Egyre több mesterséges intelligencia által generált tartalommal, szöveggel és képpel találkozunk az interneten, egyúttal pedig mind gyakrabban használjuk ezt az eszközt tanulmányaink, munkánk vagy kikapcsolódásunk során. Jelen tanulmány bemutatja a mesterséges intelligencia egyik legelterjedtebb felhasználási módjának, a nagy nyelvi modelleknek a történetét, működését és piaci jelentőségét, valamint ismerteti a technológia geopolitikai kockázatait, különösképpen a nyelv szempontjából. Ilyen kockázatok között említendő többek között a mesterséges intelligencia tanítására használt szövegek ellenőrizetlensége, nyelvi szempontú homogenitása, bizonyos kultúrkörök előnyben részesítése, valamint az így tanított és ennek nyomán bizonyos nyelveket és világnézeteket előnyben részesítő mesterséges intelligencia hatása a felhasználók gondolkodására, viselkedésére.

Kulcsszavak: mesterséges intelligencia, nagy nyelvi modellek, nyelvpolitika, geopolitika, jog

Artificial intelligence is one of the most promising technologies of our time, but also one that poses many risks. In recent years, we have ceded more and more of the control and organisation of our lives to this technology. There is an increasing number of AI-generated content, text and images on the Internet, and we are using it more and more often for study, work and leisure. This paper presents the history, functioning and market relevance of one of the most common types of AI, the large language models, and describes the geopolitical risks of the technology, with a particular focus on language preservation and maintenance. Such risks include the uncontrolled nature of the texts used to train AI, their linguistic homogeneity, the preference for certain cultures, and the impact on the

thinking and behaviour of users of AIs taught in this way, which favour certain languages and worldviews.

Keywords: artificial intelligence, large language models, language policy, geopolitics, law

1. Bevezetés

A mesterséges intelligencia (angolul AI, magyarul MI) az elmúlt évek egyik legtöbb lehetőséggel kecsegtető technológiai fejlesztése, amely komoly hatással van hétköznapjainkra, munkavégzési szokásainkra és társadalmi folyamatainkra. Az AI és annak egy alkalmazása, az úgynevezett nagy nyelvi modellek (Large Language Models, LLM) alapjaiban változtatták meg az emberi nyelvek digitális feldolgozásának folyamatát. Számos LLM-projekt, így például az OpenAI által fejlesztett ChatGPT a felhasználóbarát felülete és működése révén gyorsan elterjedt és sokak által használt alkalmazássá vált. A jelen tanulmány keretében elsőként megvizsgáljuk a nagy nyelvi modellek működésének alapjait, a modellek fejlesztéséhez elengedhetetlen feltételeket, valamint az LLM-ek társadalomra gyakorolt hatásait. Ezek ismeretében érthetjük meg azokat a problémákat, amelyeket ez az új technológia felvet, és kereshetünk jogalkotói választ e problémák megoldására.

2. A nagy nyelvi modellek fejlesztésének története

A mesterséges intelligenciát egyes kutatók a „legemberibb technológiának" nevezik, amelynek célja az emberi viselkedés és gondolkodás utánzása (Fan 2020: 8). A mesterségesen létrehozott, gondolkodó lények motívuma már az ókor világában is jelen volt, gondoljunk a görög mondakörből ismert bronzrobot, Talosz történetére (Mayor 2018; Mayor 2020: 18), az emberi gondolkodás és logika elemzése pedig már Arisztotelész kutatásainak is kiemelt témája volt (Gervai–Trautmann 2003: 39). A mai fogalmaink szerint ismert mesterséges intelligencia fejlesztéséhez szükséges tudományos, elméleti háttér lefektetése a 19. század első felében kezdődött meg. Ez volt az az időszak, amikor elsőként próbálták meg a kutatók megfogalmazni, miért lehet különbség egyes emberek mentális képességeiben, mi az a humán intelligencia, hogyan működik az emberi elme és gondolkodás (Sir Francis Galton vagy James McKeen Cattell kutatásai), majd ezen kutatások és a technológia fejlődése nyomán a 20. században kezdődtek el az első, emberi viselkedést utánzó gépek elkészítésére vonatkozó projektek (Gyekiczky 2021). A mesterséges intelligencia fejlődésének és működésének azóta számos útját ismerjük, e dolgozat szempontjából viszont elsődlegesen az úgynevezett nagy nyelvi modellek kerülnek a középpontba, amelyek olyan, nagy mennyiségű szöveg feldolgozásával tanított mesterségesintelligencia-rendszerek, amelyek képesek megérteni és létrehozni természetes nyelven szöveget (Microsoft é. n.).

A nagy nyelvi modellek fejlesztésének gyökerei egészen az 1930-as évekre nyúlnak vissza, amikor a mesterséges intelligencia kutatásának nagy úttörője, Alan Turing

létrehozta az első tárolt programú számítógépeket, emberi viselkedést és gondolkodást utánzó algoritmusokat, valamint lefektette a mesterséges intelligencia fejlesztésének alapjait (Turing 1936). Turing célja olyan gépek létrehozása volt, amelyek képesek tanulni, és a tanulás érdekében szükség esetén megváltoztatni saját belső utasításukat, kódjukat (Copeland 2025). Turing a gépek viselkedésének elemzésére és az intelligencia nyomainak keresésére létrehozta az úgynevezett Turing-tesztet, amely során egy humán felhasználónak kellett egy emberi és egy gépi partnerrel társalognia oly módon, hogy nem tudta, aktuális beszélgetőpartnere melyik. A teszt során az emberi bírálónak egy kérdéssoron végigmenve kellett eldöntenie, melyik beszélgetőpartnere volt ember, és melyik gép (Prószéky 2018: 18). Turing elmélete szerint amennyiben a gép olyan jól tud „elbeszélgetni" az emberrel, hogy képes azt megtéveszteni, akkor azt a gépet intelligensnek lehet nevezni (Fan 2020: 25). Ezt a módszertant azóta már több kritika is érte, viszont ebből a példából is láthatjuk a nyelv és a (mesterséges) intelligencia kutatásának szoros kapcsolatát.

A mesterséges intelligencia nyelvészeti felhasználásának első fontos projektjei a gépi fordító algoritmusok (machine translation, MT) voltak, amelyeket kezdetben a második világháború során rejtjelek és titkos kódok feltörése érdekében létrehozott eszközök mintájára kezdtek el fejleszteni (Cabello 2023). Az 1950-es évek hidegháborús környezetében mindkét oldal olyan eszközök fejlesztésébe kezdett, amelyek hatékonyan képesek lefordítani a más nemzetek nyelvén íródott kutatásokat, jelentéseket és egyéb dokumentumokat (Manning 2022). E kezdeti projekteket később számos, a nyelv használatához kapcsolódó kutatás követett. A természetes nyelvfeldolgozás egyik fontos mérföldköve volt a Joseph Weizenbaum által létrehozott ELIZA, amely bizonyos szempontból az első csetbotok közé tartozott. A program az emberi pszichológusra hasonlító tevékenységet végzett, és folyamatos beszélgetést volt képes folytatni felhasználóival (Fan 2020: 28; Natale 2018: 716–720). Az első ilyen, Weizenbaum által publikált beszélgetés az alábbiak szerint alakult (Weizenbaum 1966):

Felhasználó: *Men are all alike.*
ELIZA: *In what way?*
Felhasználó: *They're always bugging us about something or other.*
ELIZA: *Can you think of a specific example?*
Felhasználó: *Well, my boyfriend made me come here.*
ELIZA: *Your boyfriend made you come here?*
Felhasználó: *He says I'm depressed much of the time.*
ELIZA: *I am sorry to hear that you are depressed.*
Felhasználó: *It's true. I am unhappy.*
ELIZA: *Do you think coming here will help you not to be unhappy?*
Felhasználó: *I need some help, that much seems certain.*
ELIZA: *What would it mean to you if you got some help?*
Felhasználó: *Perhaps I could learn to get along with my mother.*
ELIZA: *Tell me more about your family.*

A különböző kezdeti algoritmusok jelentős eredményeket értek el az emberi nyelv használata terén, viszont a fejlesztésük során komoly nehézséget okozott a hardveres

lehetőségek korlátozottsága, az adatok tárolásának kihívása. A számítógépek memóriakapacitásának növekedésével mind több adat vált hatékonyan tárolhatóvá, rendezhetővé és kereshetővé, ami később a korpusznyelvészet alapját adhatta. A nagy tárolási és számítási kapacitású eszközök birtokában a kutatók képesek voltak nagy méretű szövegek statisztikai úton történő elemzésére, áttekintésére (Prószéky 2018: 21).

2013-tól kezdődően a *deep learning* (mélytanulás) technológiáját használva a kutatók hatékony nyelvi modelleket tudtak létrehozni, amelyek statisztikai alapon, elsősorban a vektoros térben történő elhelyezkedés és távolságok alapján voltak képesek modellezni a nyelveket (Manning 2022). 2023-tól kezdve a nagy nyelvi modellek mind több, nyilvános felhasználásra szánt változatát piacra dobták, így például az OpenAI által fejlesztett ChatGPT-t. A nagy nyelvi modellek hatalmas mennyiségű szöveget elemezve, abból tanulva képesek hatékonyan kommunikálni a felhasználóikkal és feladatokat megoldani (Alammar–Grootendorst 2024: 25–27). A tanulásra felhasznált szövegek mennyiségének megértéséhez példaként szolgálhat, hogy a GPT-3 modell tanítására 570 gigabyte szöveget használtak fel (Foster 2023: 259), egy nagy nyelvi modell válaszadási pontosságának 5%-kal történő növeléséhez pedig általában duplájára kell növelni a modell tanítására használt adatmennyiséget (Tranquillin–Lakshmanan–Tekiner 2023: 31).

A nagy nyelvi modellek működésük során az átalakítandó szövegeket kisebb részekre, úgynevezett tokenekre bontják, amelyek lehetnek szavak, szórészletek, de akár egyedi karakterek is. A tokeneket a modell számokká alakítja át, és önfigyelési mechanizmusa és neurális hálózata segítségével elvégzi a kért feladatokat (Alammar–Grootendorst 2024: 44–47; Chang 2024: 87; Microsoft é. n.).

Bár az LLM-technológiára épülő eszközöket a felhasználók gyakran a világ dolgairól kérdezik, az eszközök valójában nem ismerik a világot (nem képesek az emberi fogalmak szerint azt sem vizuálisan, sem más módon feldolgozni), pusztán a világról alkotott, rendelkezésükre bocsátott szövegekből dolgozva válaszolják meg a felhasználók kérdéseit (Prószéky 2024). Ez a működési modell egyrészről kitetté teszi őket a rendelkezésükre bocsátott szövegeknek, azok pontosságának és tartalmának, másrészről pedig meghatározza azokat a területeket, ahol hatékonyan lehet őket használni.

A nagy nyelvi modellek piaca mára az egyik legdinamikusabban fejlődő szektorrá vált, egyes becslések szerint 2025-ben 750 millió alkalmazás használ valamilyen nagy nyelvi modellt működése során, a piac méretét pedig 259,8 millió dollárosra becsülik 2030-ra (Uspenskyi 2025). Az LLM-ek felhasználószáma egyre jobban növekszik.

3. A mesterséges intelligencia elterjedésének jogi, geopolitikai és nyelvművelési kérdései

3.1. Geopolitika és nyelvi imperializmus

A nagy nyelvi modellek tanításához és fejlesztéséhez hatalmas szövegkorpuszok szükségesek, amelyek kapcsán alapvető kérdésként merülhet fel, hogy honnan származnak ezek a szövegek, mennyire tükrözik ezek az adatok a világ nyelvi és kulturális sokszínűségét, illetőleg miként ellenőrizhetjük a szövegek megfelelőségét, „tisztaságát”.

A mesterséges intelligencia tanítására használt adatok politikai és kulturális értelemben nem semlegesek, hiszen a nyelv a kultúránk, a világról alkotott képünk hordozója, a tanításra használt szövegek pedig jogi, társadalmi állásfoglalásokat, véleményeket tartalmazhatnak. A korpusz szövegeiben található előítéletek, torzítások és negatív megállapítások ennek nyomán könnyen hatással lehetnek a mesterséges intelligencia működésére (Alammar–Grootendorst 2024: 28; Bender–Friedman 2018: 589; Dwivedi et al. 2023: 31; Speer 2017).

Az LLM-rendszerek a világot a rendelkezésükre bocsátott szövegeken keresztül ismerik meg, ezért geopolitikai és kulturális kérdés is egyben, hogy milyen szövegekből tanulnak a ma a piacon elérhető nagy nyelvi modellek, milyen „képet" kapnak a világról, és a tanulás során előnyben részesítenek-e a fejlesztők bizonyos nyelveket, kultúrköröket.

E tanulmány fogalomhasználatában a nyelvi adatpolitika azokat a szabályozási módszereket és politikai eszközöket jelöli, amelyek meghatározzák, miként kezeljük a nyelvünkhöz kötődő adatokat, miként gyűjtjük, elemezzük és használjuk fel azokat, például a mesterséges intelligencia oktatásához. A nagy nyelvi modelleket fejlesztő vállalatok a mesterséges intelligencia tanításához szükséges adatokat legfőképpen az internetről gyűjtik (OpenAI é. n.), sok esetben nyilvános szövegtárakból, sokszor felhasználók és más cégek által létrehozott felületekről, adatbázisokból, gyakran a jogtulajdonos tudta nélkül. Egyes sajtóhírek szerint a fejlesztők egymástól átvett adatokat is felhasználnak saját modelljeik javítása érdekében (Bass–Ghaffary 2025; Foster 2023: 412).

Az így gyűjtött adatok áttekintése során kérdésként merül fel, hogy egy olyan kis nyelv, mint a magyar, hogyan lehet képes hatékony reprezentációra szert tenni a tanulásra használt szövegek körében, miként lesz képes a fejlesztés alatt álló mesterséges intelligencia hatékonyan megismerni a magyar nyelvet, ezáltal a magyar kultúrát. Már pusztán az interneten elérhető tartalmak vizsgálata során is láthatjuk a nagy nyelvek dominanciáját. Így például a GPT-1 tanításához nagyjából 7 ezer könyv és a 43%-ban angol nyelvű adatokat tartalmazó Common Crawl adatbázis szolgált alapul (Alammar–Grootendorst 2024: 20; Common Craw é. n.), a ChatGPT 2023-ban elérhető verziójának tanítására használt weboldalak 56%-a pedig angol nyelvű volt (Szalay 2023). Az angol nyelv dominanciája mindezeken túl az LLM-ek működési mechanikájában is érzékelhető, így például a modellek tokenizációs folyamatát az angol nyelvre optimalizálták, a magyarhoz hasonló agglutináló nyelvek szövegét ennek nyomán az LLM-nek több tokenre kell osztania, azaz nagyobb költségekkel lehet képes azt feldolgozni (Berryman–Ziegler 2025: 28–30). Ez a különbség nem csak a szöveggenerálásban érzékelhető: a ChatGPT beszédfelismerő és hanggeneráló funkciója angol nyelven képes a leghatékonyabban dolgozni, a szöveget beszéddé alakító modell angol hangok esetében adja a legtermészetesebb eredményt (Caelen–Blete 2024: 21).

A nagy nyelvek AI-fejlesztésben betöltött szerepe komoly kockázatot jelent, és egyfajta „nyelvi imperializmushoz" vezethet, hiszen a technológia működését, „gondolkodását", „világról alkotott képét" a tanításhoz használt szövegeken keresztül néhány nagy nyelv és az azokhoz kötődő országok határozzák meg.

Az Európai Parlament a technológia fejlődésében rejlő lehetőségeket, valamint az itt bemutatott kockázatokat már 2018-ban felismerte akkor, amikor kiadta

állásfoglalását a nyelvek közötti egyenlőségről a digitális korban (Európai Parlament 2018). Az állásfoglalás kiemeli, a digitális technológia az elmúlt évtizedekben komoly hatással volt a nyelvek fejlődésére, megjegyzi továbbá, hogy a „kevésbé használt európai nyelvek jelentős hátrányba kerülnek az eszközök, források és a kutatási finanszírozás akut hiánya miatt, amely akadályozza a kutatók tevékenységét, és szűkíti a kutatásaik körét, és még abban az esetben is, ha a kutatók rendelkeznek a szükséges technológiai készségekkel, nem képesek maximálisan kiaknázni a nyelvi technológiákat" (az állásfoglalás 2. és 3. bekezdése). Az Európai Parlament álláspontja szerint egyre komolyabb a digitális szakadék a széles körben használt és a kevésbé használt nyelvek között, egyúttal pedig megjegyzi, hogy az egyes tartalmak különböző nyelveken történő elérhetővé tétele hozzájárulhatna a nyelvek közötti esélyegyenlőség növeléséhez (az állásfoglalás 4. bekezdése). Mivel az egyes vállalatok elsődlegesen angol nyelven teszik elérhetővé szolgáltatásaikat, ezen a nyelven történik főként a fejlesztés, ezért elengedhetetlen az egységes fellépés annak érdekében, hogy a nyelvi technológiákat a kevésbé használt uniós nyelveken is elérhetővé tegyék (az állásfoglalás 6. bekezdése).

Az Európai Parlament 2021-ben kiadott állásfoglalásában (Európai Parlament 2021) felhívja a figyelmet a mesterséges intelligencia által generált kockázatokra a nyelvi sokszínűségre nézve, egyúttal viszont pozitívumként említi, hogy a technológia megfelelő alkalmazás mellett segítheti a többnyelvűséget (az állásfoglalás AH. bekezdése). Az állásfoglalás kiemeli, az Európai Unióban heterogén környezetet kell létrehozni a nyelvi sokszínűség fenntartása érdekében (az állásfoglalás 63. pontja). A digitális nyelvi esélyegyenlőség érdekében az Európai Unió támogatja a European Language Equality (ELE) projektet, amely célja a komplex cselekvési terv kidolgozása, a természetesnyelv-feldolgozással kapcsolatos erőforrások felmérése és elemzése (Jelencsik–Mátyus–Héja–Varga–Váradi 2022: 1). Láthatjuk tehát, hogy az Európai Unió azon az álláspontón van, hogy ha szükséges, akár jogalkotói eszközökkel is hozzá kell járulnunk a nyelvek esélyegyenlőségének növeléséhez.

Mindezen kockázatok mellett viszont a nagy nyelvi modellek számos pozitívumot is hordoznak magukban. Ezen alkalmazások oktatásban, hatékonyságnövelésben betöltött szerepe vitathatatlan, ahogy a nyelvek és kultúra megőrzésének is fontos eszközeivé válhatnak. Prószéky Gábor egy interjújában felhívja a figyelmet arra, hogy a mesterséges intelligencia segítségünkre lehet a kihalás szélén álló nyelvek megőrzésében, hozzájárulhat ezen nyelvek megmaradásához (Magyar Hírlap 2024). A nagy nyelvi modellek mindemellett hatékonyak lehetnek az egyes szolgáltatások és az azokhoz kapcsolódó ügyféltámogatás több nyelven történő elérhetőségének biztosításában. Az LLM-ek segítségével hatékonyan, kis költségráfordítással tudunk az egyes felhasználóknak saját nyelvükön ügyfélszolgálatot biztosítani.

A nagy nyelvi modellek és a geopolitika kapcsolatának elemzése során fontos kiemelni, a jelenleg a piacon elérhető eszközök többnyire zárt szolgáltatások, amelyekhez tipikusan csak internetes API-hozzáférést kap a felhasználó, így belső működésük nem nyilvános. Ez egyrészről előny is, hiszen nem szükséges nagy számítási kapacitású számítógép vagy szerver a hozzáféréshez, másrészről viszont azt is jelenti, hogy saját célra csak nehezen finomhangolható egy ilyen eszköz, valamint (akárcsak a legtöbb platformszolgáltatás esetében) a felhasználók teljes egészében a szolgáltató infrastruktúrájától függenek, ami tovább növelheti a kisebb országok kitettségét

a nagyobb országok és azok vállalatai felé, és hozzájárul az egyes felhasználók adatainak külföldre kerüléséhez (Alammar–Grootendorst 2024: 30).

3.2. A nagy nyelvi modellek hatása a gondolkodásra, a világról alkotott képre

A nagy nyelvi modellekre épülő technológiák, mint a ChatGPT, természetszerűen azokra a szövegekre építik a felhasználóknak adott válaszaikat, amelyekből tanításuk történt, nem a világot magát, hanem a világról szóló szövegeket ismerik. Éppen ezért fontos kérdés, hogy a rendelkezésükre bocsátott szövegek milyen tartalommal bírnak, milyen kulturális hátteret és világnézetet tükröznek. A jogról, politikáról, etikáról alkotott kép országonként, kultúránként eltérő. Ami az egyik ország közéletében, közbeszédében bevett és elfogadott nyilatkozat, máshol tiltott lehet. Erre példa a holokauszttagadás, amely számos európai országban kívül esik a szabad véleménynyilvánítás körén, sőt több helyen – így többek között hazánkban is – bűncselekménynek minősül, az amerikai jogi kultúrában viszont védett beszédnek tekintendő (Btk. 333. §, Koltay 2019: 16–17). Az ilyen és ehhez hasonló eltérések könnyen beépülhetnek a mesterséges intelligencia által adott válaszokba, azáltal pedig, hogy a lakosság körében mind népszerűbb ez a technológia, mind gyakrabban használjuk információk gyűjtésére, könnyen aktív befolyásolójává válhat a közbeszédnek, közvélekedésnek.

A kulturális eltérések mellett fontos kiemelni, a nagy nyelvi modellek képesek meggyőzően hangzó, viszont tényszerűen hamis vagy ok-okozati összefüggést nélkülöző válaszokat adni felhasználóiknak, gyakran még hamis hivatkozásokat is generálva. A felhasználók többsége ezeket a hibás válaszokat nem feltétlenül ismeri fel, így azok beépülhetnek gondolatvilágukba, világról alkotott képükbe (Alkaissi–McFarlane 2023: 4; Cheng–Calhoun–Reedy 2025: 2–3; Foster 2023: 411–413).

Az AI viszont nemcsak tartalmát tekintve befolyásolhatja az emberek vélekedését a világról, hanem hatással lehet nyelvhasználatunkra, kommunikációs szokásainkra is, egyrészt azáltal, hogy a felhasználók átveszik az AI által generált szövegek stílusát, másrészről pedig beépítik ezeket a szövegeket iskolai beadandóikba, az általuk írt cikkekbe, levelekbe, bejegyzésekbe. Ezáltal a mesterséges intelligencia formálójává válhat a nyelvnek, és hozzájárulhat más, nagyobb nyelvekben használt kifejezések elterjedéséhez.

4. Mit tehet a jogalkotó, hogy a mesterséges intelligencia jól beszéljen magyarul?

A magyar nyelv digitális jövője elsődlegesen a nyelvi adatpolitikánkon, azaz azon múlik, hogy miként gyűjtjük, kezeljük nyelvi adatainkat, és tesszük azokat hozzáférhetővé a mesterséges intelligenciát fejlesztő cégeknek.

Ezen törekvés fontos eszközei az állam által támogatott, professzionális csapat által szerkesztett, jól strukturált és gondozott szövegadatbázisok, amelyek – amennyiben azok korlátozás nélkül hozzáférhetők a fejlesztő cégek számára – alapjául szolgálhatnak a mesterséges intelligencia magyar nyelven történő tanításához.

Hazánkban ilyen hozzáférhető szövegadatbázis a Magyar Nemzeti Szövegtár, amelynek fejlesztése 1998 elején kezdődött, adatbázisa pedig jelenleg 187,6 millió szövegszót tartalmaz. Szintén fontos megemlíteni a Szeged Korpuszt, valamint a Történeti Korpuszt. Utóbbi 1772 és 2000 között létrejött szövegek értékes tárát adja a kutatók számára (Prószéky–Olaszy–Váradi 2006). Álláspontom szerint rendkívül fontos a mesterséges intelligencia magyar nyelvű tanításához az, hogy minél több, akár államilag támogatott nyilvános szövegkorpusz jöjjön létre tudományterületekre történő bontásban. Ez hozzájárulna ahhoz, hogy a fejlesztők hiteles, ellenőrzött adatbázisokból tudják tanítani algoritmusaikat, az ingyenes hozzáférés pedig vonzóbbá tenné ezek alkalmazását más, költségesebb alternatívákhoz képest. A hazai példákon felül fontos kiemelni, számos nemzetközi adatbázis tartalmaz magyar nyelvű alkorpuszt, így például az OPUS vagy a CCMatrix (Jelencsik–Mátyus–Héja–Varga–Váradi 2022: 2).

A mesterséges intelligencia által generált szövegek mindinkább beépülnek életünkbe. Számos szakma képviselője használja a mesterséges intelligenciát és generál vele különböző tartalmakat munkája során, például több marketingvállalat alkalmaz AI generálta tartalmakat (Foster 2023: 409; Kiss 2024). A ChatGPT-hez hasonló szolgáltatások felhasználói a mesterséges intelligencia által generált tartalmakat felhasználva dolgoznak, tanulnak, értik meg a világot. Éppen ezért elengedhetetlen, hogy egyértelmű különbséget tegyünk az ember és gép által létrehozott szövegek között, azokat akár címkézéssel, akár más módon elkülönítsük egymástól annak érdekében, hogy a felhasználó, azaz az információ után kutató állampolgár egyértelműen tájékoztatva legyen arról, hogy az a szöveg, amit felhasznál, a mesterséges intelligencia munkája-e. Ez a mesterséges intelligencia tanítása szempontjából is fontos információ lehet, hiszen így elkerülhető, hogy AI által generált szövegek kerüljenek be a mesterséges intelligencia tanítására használt szövegtárba.

A mesterséges intelligencia által generált szövegek felhasználásához kapcsolódóan a másik fontos állami feladat a technológia alapjainak a köznevelésbe és a felsőoktatásba történő beépítése, annak érdekében, hogy a felhasználók megértsék a nagy nyelvi modellek korlátait, egyúttal pedig felelősen használják ezeket az eszközöket a hétköznapjaikban. Álláspontom szerint elengedhetetlen, hogy legalább a középiskolai oktatásban hangsúlyos szerepet kapjon az AI tanulásban, munkavégzésben történő alkalmazhatóságának bemutatása, ami hozzájárulhat a diákok, a jövőbeli munkavállalók hatékonyságához. Fontos, hogy a technológia működésének ismertetése révén megértessük a következő generációkkal, milyen kockázatokat rejt magában a mesterséges intelligenciába vetett feltétlen bizalom, a technológia felelőtlen használata.

A mesterséges intelligencia komoly segítséget jelenthet a diákok számára, például a személyre szabott tutorálás keretében, viszont komoly kihívások elé állítja a klasszikus oktatási rendszerünket. Több szerző felveti kérdésként, mennyire lesz hatékony a jövőben egy esszéfeladat, ha a diák már a mobiltelefonján is képes magas minőségű szövegeket generálni. Az AI generálta szövegek nyomán a plágium fogalma kibővülhet, ami egyrészről lépésre kényszerítheti a jogalkotót is (Foster 2023: 410–411), másrészről pedig kihívás elé állíthatja a készségfejlesztéssel foglalkozó szakembereket, oktatókat, akiknek meg kell értetniük a diákokkal, hogy a beadandó feladat önálló megírása az ő készségeik fejlesztését szolgálja, az otthoni esszé pedig a nyelv használatának hatékony elsajátításában, a tananyag megértésében segít.

Az oktatásba történő beépítés mellett elengedhetetlen, hogy az idősebb generációk számára is hatékony tájékoztatást és felkészítést biztosítsunk, hiszen a mesterségesen generált tartalmak könnyen a csalók eszközévé válhatnak a technológiát kevésbé ismerő korcsoportok megtévesztése során. A lakosság AI-tudatosítására a korábbi években több állami törekvéssel is találkozhattunk, így például a Digitális Jólét Programmal, amelynek kimondott célja volt, hogy az állampolgárok a technológiai forradalom nyertesei lehessenek; 2016-ban pedig Magyarország kormánya elfogadta Magyarország Digitális Oktatási Stratégiáját, amely szintén fontos alapja lehet a mesterséges intelligencia oktatásba történő hatékonyabb beépítésének (Tóth 2022: 41–48).

A nyelvünk és kultúránk megőrzésének fontos eszköze lehet továbbá a saját, hazai nagy nyelvi modellek készítése és ezen fejlesztések támogatása. Erre pozitív példa lehet a PULI, a Nyelvtudományi Kutatóközpont fejlesztése, amely 50 milliárd szavas magyar szövegkorpusszal rendelkezik. Az ilyen és ehhez hasonló projektek állami támogatásával hozzájárulhatunk a magyar nyelvet jól ismerő és használó mesterségesintelligencia-alapú alkalmazások elterjedéséhez, egyúttal pedig a technológia hazai fejlesztéséhez.

A magyar nyelv támogatásában nagy szerepet játszik a fogyasztóvédelmi jogi szabályozás és az ehhez kapcsolódó hatósági fellépés. A Gazdasági Versenyhivatal (GVH) az európai versenyhatóságok közül elsőként végzett átfogó piacelemzést, és állapította meg, hogy a nagy vállalatok üzletpolitikájának nem része a kis nyelvekre alapozott öntanuló rendszerek fejlesztése. A GVH egyúttal felhívta a figyelmet az állami beavatkozás fontosságára is (Gazdasági Versenyhivatal 2024). A Hatóság 2023-ban eljárást indított a Microsoft európai leányvállalatával szemben a Bing felületébe integrált nagy nyelvi modell miatt (Gazdasági Versenyhivatal 2023), amely eljárás nyomán a Microsoft kötelezettséget vállalt arra, hogy egy legalább 10 milliárd magyar szót tartalmazó adatbázist hoz létre, amely alapul szolgálhat LLM-jének magyar nyelven történő tanításában, és az adatállományt a vállalat más piaci szereplők számára is elérhetővé teszi (Gazdasági Versenyhivatal 2025).

A kis nyelvek támogatása sokak szerint nem valósítható meg pusztán jogalkotói beavatkozás segítségével, elengedhetetlen az, hogy a nagy nyelvi modellek fejlesztői önszabályozás, valamint működési struktúrájuk áttekintése révén felelősséget vállaljanak az általuk üzemeltetett modell nyelvi és kulturális sokszínűségéért. Ennek eszköze lehet a nyelvi tekintetben diverz fejlesztői csapat létrehozása, amelynek tagjai saját anyanyelvükön lehetnek képesek hatékonyan tesztelni a modell működését, és felismerni annak esetleges hiányosságait vagy hibáit (Alammar–Grootendorst 2024: 377). Mindemellett fontos, hogy a fejlesztő vállalatok külön figyelmet fordítsanak az általuk készített modellek kis nyelvekre történő finomhangolására, ahogy teszi ezt például a GitHub a Codex modelljével (Berryman–Ziegler 2025: 74).

5. Összegzés

A nagy nyelvi modellek számos lehetőséget rejtenek magukban az oktatás, munkavégzés és tartalomgyártás területén, egyúttal viszont, mint minden technológia, komoly

kockázatot is jelenthet felelőtlen használatuk, elterjedésük pedig kihívás elé állíthatja a jogalkotót is, akinek egyrészről feladata és célja ösztönözni az innovációt, a piaci hatékonyság növelését, másrészről viszont kötelessége hozzájárulni kulturális értékeink megőrzéséhez, terjesztéséhez. A nyelvi adatok megfelelő rendezése és hozzáférhetővé tétele lehetőséget biztosíthat arra, hogy a jogalkotó mindkét célt hatékonyan tudja szolgálni, ahogy a mesterséges intelligencia oktatásba történő beépítése is, ami egyszerre teheti versenyképessé a következő generáció munkavállalóit, és segíthet az AI felelős használatában. Ehhez viszont elengedhetetlen az aktív állami támogatás mind a források biztosítása, mind pedig az oktatás átalakítása révén.

Szakirodalom

Alammar, Jay – Grootendorst, Maarten 2024: *Hands-On Large Language Models – Language Understanding and Generation.* Sebastopol: O'Reilly Media.

Alkaissi, Hussam – McFarlane, Samy I. 2023: Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 15/2: e35179. Online: https://doi.org/10.7759/cureus.35179

Bass, Dina – Ghaffary Shirin 2025: Microsoft Probing If DeepSeek-Linked Group Improperly Obtained OpenAI Data. *Bloomberg,* 2025. január 29. Online: https://www.bloomberg.com/news/articles/2025-01-29/microsoft-probing-if-deepseek-linked-group-improperly-obtained-openai-data?embedded-checkout=true

Bender, Emily M. – Friedman, Batya 2018: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6: 587–604. Online: https://doi.org/10.1162/tacl_a_00041

Berryman, John – Ziegler, Albert 2025: *Prompt Engineering for LLMs – The Art and Science of Building Large Language Model–Based Applications.* Sebastopol: O'Reilly Media.

Cabello, Adria 2023: The Evolution of Language Models: A Journey Through Time. *Medium,* 2023. november 2. Online: https://medium.com/@adria.cabello/the-evolution-of-language-models-a-journey-through-time-3179f72ae7eb

Caelen, Olivier – Blete, Marie-Alice 2024: *Developing Apps with GPT-4 and ChatGPT – Build Intelligent Chatbots, Content Generators, and More*. Sebastopol: O'Reilly Media.

Chang, Susan Shu 2024: *Machine Learning Interviews Kickstart Your Machine Learning and Data Career.* Sebastopol: O'Reilly Media.

Cheng, Adam – Calhoun, Aaron – Reedy, Gabriel 2025: Artificial Intelligence-Assisted Academic Writing: Recommendations for Ethical Use. *Advances in Simulation* 10/22. Online: https://doi.org/10.1186/s41077-025-00350-6

Common Craw [é. n.]: *Statistics of Common Crawl Monthly Archives.* Online: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Copeland, B. J. 2025: History of Artificial Intelligence. *Encyclopedia Britannica,* 2025. február 25. Online: https://www.britannica.com/science/history-of-artificial-intelligence

Dwivedi, Yogesh K. et al. 2023: "So What if ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management* 71: 102642. Online: https://doi.org/10.1016/j.ijinfomgt.2023.102642

Fan, Shelly 2020: *Lecserél-e minket a mesterséges intelligencia?* Budapest: Scolar Kiadó.

Foster, David 2023: *Generative Deep Learning – Teaching Machines to Paint, Write, Compose, and Play.* Sebastopol: O'Reilly Media.

Gazdasági Versenyhivatal 2023: A GVH megvizsgálja, hogy a Microsoft miként tájékoztatja a fogyasztókat új keresőszolgáltatásáról. *Gazdasági Versenyhivatal*, 2023. július 25. Online: https://gvh.hu/sajtoszoba/sajtokozlemenyek/2023-as-sajtokozlemenyek/a-gvh-megvizsgalja-hogy-a-microsoft-mikent-tajekoztatja-a-fogyasztokat-uj-keresoszolgaltatasarol

Gervai Pál – Trautmann László 2003: Filozófiai irányzatok és jövőképek – Az Ókor. In: Nováky Erzsébet – Kristóf Tamás: *Ókori jövőképek és jövőkutatás a vállalati gyakorlatban.* Budapest: Budapesti Közgazdaságtudományi és Államigazgatási Egyetem Jövőkutatási Kutatóközpont.

Gyekiczky Tamás 2021: A digitális társadalom olvasatai – VII. *Jogászvilág*, 2021. január 7. Online: https://jogaszvilag.hu/vilagjogasz/a-digitalis-tarsadalom-olvasatai-vii/

Jelencsik-Mátyus Kinga – Héja Enikő – Varga Zsófia – Váradi Tamás 2022: D1.18 Report on the Hungarian Language. *European Language Equality*, 2022. február 28. Online: https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_18__Language_Report_Hungarian.pdf

Kiss Tamás 2024: A jövő marketingosztálya: az AI és emberi vezetés harmóniája. *Kontext*, 2024. szeptember 16. Online: https://bit.ly/4iHG1Dc

Koltay András 2019: A social media platformok jogi státusa a szólásszabadság nézőpontjából. In: *In Medias Res* 8/1: 1–56.

Manning, Christopher D. 2022: Human Language Understanding & Reasoning. *Daedalus* 151/2: 127–138. Online: https://doi.org/10.1162/daed_a_01905

Mayor, Adrienne 2018: Tyrants and Robots in Ancient Greece. *History Today*, 2018. november 11. Online: https://www.historytoday.com/archive/feature/tyrants-and-robots-ancient-greece

Mayor, Adrienne 2020: *Gods and Robots Myths, Machines, and Ancient Dreams of Technology.* Princeton: Princeton University Press.

Natale, Simone 2018: If Software Is Narrative: Joseph Weizenbaum, Artificial Intelligence and the Biographie of ELIZA. *New Media & Society* 21/3: 712–728. Online: https://doi.org/10.1177/1461444818804980

Prószéky Gábor – Olaszy Gábor – Váradi Tamás 2006: Nyelvtechnológia. In: Kiefer Ferenc (szerk.): *Magyar nyelv.* Budapest: Akadémiai Kiadó.

Prószéky Gábor 2018: A nyelvtudomány találkozásai a gépi intelligenciával In: Tolcsvai Nagy Gábor (szerk.): *A humán tudományok és a gépi intelligencia*. Budapest: Gondolat Kiadó.

Szalay Klára 2023: Mesterséges intelligencia, szabályozási irányok. Elemzés Országgyűlési képviselők részére. *Parlament,* 2023. szeptember 20. Online: https://www.parlament.hu/documents/d/guest/mesterseges-intelligencia

Tóth András 2022: *A digitális állam információbiztonsági kihívásai*. Budapest: Ludovika Egyetemi Kiadó.
Tranquillin, Marco – Lakshmanan, Valliappa – Tekiner, Firat 2023: *Architecting Data and Machine Learning Platforms – Enable Analytics and AI-Driven Innovation in the Cloud.* Sebastopol: O'Reilly Media.
Turing, Alan 1936: On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 2/42: 230–267. Online: https://doi.org/10.1112/plms/s2-42.1.230
Uspenskyi, Serhii 2025: Large Language Model Statistics and Numbers. *Springs,* 2025. február 21. Online: https://springsapps.com/knowledge/large-language-model-statistics-and-numbers-2024
Weizenbaum, Joseph 1966: ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9/1: 36–45. Online: https://doi.org/10.1145/365153.365168

Források

A Büntető Törvénykönyvről szóló 2012. évi C. törvény (Röviden: Btk.)
Európai Parlament 2018: *Az Európai Parlament 2018. szeptember 11-i állásfoglalása a nyelvek közötti egyenlőségről a digitális korban* (2018/2028(INI)) (2019/C 433/07)
Európai Parlament 2021: *Mesterséges intelligencia az oktatásban, a kulturális és az audiovizuális ágazatban – Az Európai Parlament 2021. május 19-i állásfoglalása a mesterséges intelligencia alkalmazásáról az oktatásban, a kulturális és az audiovizuális ágazatban* (2020/2017(INI)) (2022/C 15/04)
Gazdasági Versenyhivatal 2024: A mesterséges intelligencia piaci versenyre és fogyasztókra gyakorolt hatásainak vizsgálata – A Gazdasági Versenyhivatal AL/234/2024. számú piacelemzésének eredményeként készült tanulmány. *Gazdasági Versenyhivatal,* 2024. október 21. Online: https://gvh.hu/pfile/file?path=/dontesek/agazati_vizsgalatok_piacelemzesek/piacelemzesek/mesterseges-intelligencia_piacelemzes_tanulmany_2024_10_21&inline=true
Gazdasági Versenyhivatal 2025: Magyarul tanítja a Microsoft a mesterséges intelligenciát a GVH-nak köszönhetően. *Gazdasági Versenyhivatal,* 2025. május 30. Online: https://gvh.hu/sajtoszoba/sajtokozlemenyek/2025-os-sajtokozlemenyek/magyarul-tanitja-a-microsoft-a-mesterseges-intelligenciat-a-gvh-nak-koszonhetoen
InfoRádió - Infostart [@InfoRádió-Infostart]: Veszélyben van-e a magyar nyelv? Mit hoz a digitális világ? Prószéky Gábor, InfóRádió, Aréna. *YouTube,* 2024. augusztus 22. Online: https://www.youtube.com/watch?v=WzNTXn0gZ3Y
Magyar Hírlap 2024: A mesterséges intelligencia és a nyelvhasználat közötti kapcsolat. *Magyar Hírlap*, 2024. március 27. Online: https://www.magyarhirlap.hu/tech/20240327-a-mesterseges-intelligencia-es-a-nyelvhasznalat-kozotti-kapcsolat

Microsoft [é. n.]: *Mik a nagy nyelvi modellek (LLM-ek)?* Online: https://azure.microsoft.com/hu-hu/resources/cloud-computing-dictionary/what-are-large-language-models-llms

OpenAI [é. n.]: *How ChatGPT and Our Foundation Models are Developed.* Online: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed

Speer, Robyn 2017: ConceptNet Numberbatch 17.04: Better, Less-Stereotyped Word Vectors. *ConceptNet Blog,* 2017. április 17. Online: https://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

Kolláth Mihály Gábor jogász, politológus, közgazdász, az Eötvös Loránd Tudományegyetem Állam- és Jogtudományi Karának egyetemi tanársegédje, a JÖSz Intézet főigazgatója, a Futurum Innovációs Alapítvány kutatási és fejlesztési vezetője. Kutatási területei: a mesterségesintelligencia-alapú hatékonyságnövelés, a mesterséges intelligencia szabályozása, médiajog, digitalizáció és társadalom.