https://doi.org/10.56233/SZABADPIAC.2020.1.9

ELMÉLETILEG

Koltay Tibor

A MESTERSÉGES INTELLIGENCIA ÉS AZ ADATVEZÉRELT VILÁG. ISMERETEK ÉS KÉSZSÉGEK

Bevezetés

Számos jóslat szól arról, hogy a mesterséges intelligencia korában bizonyos területeken nem lesz szükség emberi közreműködésre (Floridi, 2016, 2017), ide értve információk és adatok nagy adatállományokból történő kinyerését.

Azt is be kell látnunk, hogy newtoni világban élünk, tehát ami megtörtént, nem visszafordítható, és csak lejegyzett, rögzített manifesztációjában, vagyis információ és adat formájában tudunk majd értesülni róla (Floridi, 2015). Ez azért különösen fontos és meghatározó jelentőségű, mert a ma élő generációk egy részének tagjai az utolsók, akik még megtapasztalták, milyen a teljesen offline és analóg világ. A jövőben viszont ennek a tapasztalatnak híján szükségük lesz olyan tudásra, amelynek kevés előzménye van, mivel eddig csak egy részét sajátíthatták el (Floridi, 2018). Ennek fényében érdemes elgondolkoznunk a jelenben és a jövőben elsajátítandó készségek, képességek és jártasságok természetéről. Ezek egy része éppen a fentebb említett rögzített formákhoz, így az információhoz és az adatokhoz kötődik.

Ennek a feladatnak megfelelően, ez az írás a téma következő részkérdéseit érinti:

- az adatok szemlélete,
- a nagy adatok,
- az információs és adattúlterhelés,
- az adatok minősége,
- az adatok kritikai szemlélete.

Hogyan látjuk ma az adatok természetét?

Ahhoz, hogy közelebb kerüljünk a címben szereplő, adatvezérelt világ természetének megértéséhez, érdemes megvizsgálnunk, hogy miben változott az adatok jellemzőinek és szerepének megítélése. Ez a megítélés ugyanis sokáig kötődött ahhoz a hierarchikus szemlélethez, amelynek alapján az adatok, az információ, a tudás és a bölcsesség viszonyát próbáltuk leírni. Ez általában arra az elképzelésre épült, hogy az adatok egy piramis alján helyezkednek el, amely az adatok, az információ, a tudás és a bölcsesség viszonyrendszerét képezi le, úgy, hogy ezek meghatározásait elsősorban bizonyos jellemzők hiányára építve adja meg a következő módon:

- Az adatok dolgok, események, tevékenységek és tranzakciók elemi és rögzített leírásai.
- Az adatok diszkrét, objektív tények vagy megfigyelések, amelyek szervezetlenek és feldolgozatlanok, továbbá nem közvetítenek specifikus jelentést.
- Az adatoknak nincsen jelentése (értelme) vagy értéke, mivel nincs kontextusuk, és nem rendelkeznek interpretációkkal.

Ezzel szemben a formátum, a struktúra, a szervezettség, jelentés és érték köré szerveződve az az információ meghatározásai a következőket tartalmazzák:

- Formattált adat, amely a valóság reprezentációja.
- Olyan adat, amely egy tárgy megértéséhez többletértéket ad.
- Olyan adat, amelyet úgy formáltak/szerveztek, hogy emberi lények (a befogadók) számára értelemmel bírjon, és hasznos legyen (Rowley, 2007).

Ezeket a meghatározásokat nézve már korábban is felmerült annak szükségesége, hogy elgondolkodjunk azon, van-e éles határvonal az adat, az információ s a tudás között, vagy pedig egy kontinuumot alkotnak, amelyben a jelentés, a struktúra és a „cselekvésre alkalmas" jelleg különböző szinteken jelennek meg. Ma pedig már látjuk, hogy az adatok és az információk közötti kapcsolatot a korábbinál differenciáltabb módon kell értelmeznünk (Makani, 2015), tehát egy kevésbé a hierarchiára épülő gondolkodásmód alapján például sokkal rugalmasabb lehet a szemléletünk. Ezt támasztja alá, hogy ontológiai szempontból nézve az információ és az adatok egyaránt jelek sorozatainak, tehát egymás közeli rokonainak tekinthetők (Yu, 2015).

Ennek ellenére sokan megkérdőjelezik, hogy egyenlőségjel tehető-e az adat és az információ közé (Špiranec–Kos–George, 2019).

Azt mindenesetre elfogadhatjuk, hogy az adatok megközelítése nem támaszkodhat csupán az empirizmusra és az indukcióra, valamint az ezekre épülő hierarchikus modellekre, hanem Peirce (1960) pragmatista szemiotikáján, tehát a jelek olyan megközelítésén kellene alapulnia, amely figyelembe veszi az interpretálható tárgyak, szimbólumaik és interpretációik közötti összefüggéseket. Ezt szem előtt tartva adatnak tekinthetünk bármit, ami szemantikai és pragmatikai szempontból megfelelő módon rögzíthető adatbázisokban. Amit rögzítünk, annak szemantikai szempontból igaz vagy hamis állításnak kell lennie. A pragmatikai megközelítés pedig megköveteli, hogy konkrét tényeknek tekintsük őket (Frické, 2019).

A rögzített jelleg szükségszerűsége megjelenik Buckland (1991) sokak által elfogadott tipológiájában, amelyben az információnak három létformáját különbözteti meg. Az első létforma (az információ mint tudás) azonos az átadott tudással. A második létforma (az információ mint folyamat) tudatállapotunkat módosíthatja. A harmadik létforma (az információ mint dolog) esetében az információra úgy tekintünk, mint kézzelfogható, rögzített entitásra, amelyet ki tudunk fejezni, le tudunk írni, reprezentációk formájában tükrözni tudunk, vagy fizikailag (jelként) képviselve van.

A megváltozott szemlélet viszont nem módosította az adat, az információ és a tudás származtatásával kapcsolatos konszenzusnak azt az elemét, amelynek alapján egymásból határozhatók meg, bár az adat és az információ egyaránt szerepel a tudás inputjaként. Megerősítve látjuk azt is, hogy a három fogalom között hierarchikus a kapcsolat, még akkor is, ha az nem kizárólagos. Nincs viszont egyetértés azoknak a folyamatoknak a természetéről, amelyek az adatokat, az információt és a tudást összekötik.

Ha viszont a tudáspiramis kapcsán arról a lépésről beszélünk, amikor az adatokból információ lesz, nem vesszük figyelembe a megfigyelhető és az elméleti jelenségek közötti különbséget, bár a tudomány az utóbbinak több esetét is ismeri (Frické, 2019).

Ha az adatoknak azt a kiemelkedő fontosságú csoportját tekintjük, amelyet kutatási adatoknak nevezünk, világosan látható, hogy azok nemcsak empirikus kutatómunka eredményei, vagy

statisztikai elemzések nyersanyagai, hanem „saját jogon" kutatási tárgyak is (Pryor, 2012). Ha pedig azt nézzük, hogy az adatokat egészen az 1990-es évekig nem értékes üzleti forrásnak, hanem olyan melléktermékneknek tekintették, amely a kereskedelmi tranzakciók befejeztével elveszti értéket, jól láthatjuk az ezzel kapcsolatos szemlélet változását, hiszen éppen az üzleti világ az, ahol ma már nagy jelentőséget tulajdonítanak az adatoknak és (különösen) azok minőségének (Al-Ruithe–Benkhelifa–Hameed, 2018). Természetesen tudatában kell lennünk annak is, hogy mindkét megközelítés egy-egy viszonylag szűk szakmai közösség számára jelzi az adatok fontosságát (Pappas–Emmelhainz–Seale, 2016).

Problémák a nagy adatok körül

Köztudottnak tekinthető az a tény, hogy a mesterséges intelligencia jelentős mértékben a nagy adatokra (Big Data) épül. A nagy adatok viszont egyaránt támaszkodnak a technológia erejére, a hatékony elemzés lehetőségére és a mitológiára. Az utóbbi alapja az a hit, hogy a nagy adatok magasabb rendű intelligenciát jelentenek, olyan tudást felmutatva, amelyre egyaránt jellemző az igazság, az objektivitás és a pontosság. A nagy adatok objektivitását azért is nehéz megítélni, mert a velük kapcsolatos diskurzusnak egyaránt része az utópia és a disztópia (anti-utópia). Boyd és Crawford (2012) szerint az előbbi példája az a mítosz, amely szerint ezek az adatállományok a tudás és az intelligencia eddiginél fejlettebb formáját kínálják, olyan meglátásokat lehetővé téve, amelyek eddig nem voltak elképzelhetők. Ehhez hozzáteszik, hogy bár a nagy adatok számos társadalmi jelenséget tesznek számszerűsíthetővé, továbbra is szubjektívek maradnak, viszont amit számszerűsítenek, az nem feltétlenül kerül közelebb az objektív igazsághoz.

A másik oldalon, megint csak Boyd és Crawford (2012) említik azokat a félelmeket, amelyek annak lehetőségét villantják fel, hogy a nagy adatok segítségével lehetővé válik a magánélet titkosságának megsértése, a szabadságjogok csorbítása, az állam és a cégek által gyakorolt ellenőrzés növelése. Ezek az ellentmondások a mesterséges intelligenciára is igazak kell, hogy legyenek.

Az információs és adattúlterhelés

Nem véletlen, hogy sok kutató jobban szeret kisebb, „közelről is szemügyre vehető" adatmennyiségekkel dolgozni. Ez arra vezethető vissza, hogy az adatok mennyiségének növekedésével fordítottan arányos az egyes egyedeket megfigyelő képességünk, ráadásul nehezebb megállapítani, hogy ezek a megfigyelések mit jelentenek, mivel nem feltétlenül tudjuk, hogy miként gyűjtötték, hogyan kezelték vagy alakították át őket (Borgman, 2015).

Az adatok nagy mennyisége további gondokat is okozhat. Floridi (2012) szerint például nem csupán az a probléma, hogy milyen mennyiségű adatot tudunk technológiai eszközökkel kezelni, hanem az is, hogy a hatékonyabb technikák és technológiák segítségével a számítógépek minden másnál nagyobb mértékű túlterhelést okoznak. A túlterhelés fő

jellemzője, hogy akadályozza az információk értelmezését, ezzel zajjá válik, amely szennyezi a környezetét (Morville 2005).

Természetesen azt, hogy az információfeldolgozás iránti igény meghaladja a feldolgozásra fordítható időt és erőfeszítést (Pijpers 2010), nemcsak az üzenetek túl nagy száma okozhatja, hanem az is, hogy a bejövő üzenetek nincsenek megfelelően szervezve ahhoz, hogy könnyen felismerhető legyen, melyik a fontos közülük (Jones–Ravid–Rafaeli, 2004).

Korábban többnyire csak az információs túlterhelés – egyébként igen összetett – kérdései voltak ismertek. Az üzenetek túl nagy száma, amit az információs túlterhelés kapcsán a TMI (Too Much Information, túl sok információ) betűszóval szoktunk illetni, a probléma mennyiségi oldalát mutatja meg, másképpen szólva makroszintjének jellemzője. Ami viszont az információs túlterhelés minőségi (kvalitatív), azaz mikroszintjéhez köthető, az az a tény, hogy az információ sokféleségével kell szembenéznünk, miközben nem rendelkezünk megfelelő szűrőkkel a használható információk kiválasztására, vagy nem megfelelően használjuk ezeket a szűrőket. Mindez arra vezethető vissza, hogy sokakban nincsenek meg a szűrést megalapozó kritikai gondolkodáshoz szükséges beállítódások és készségek, ami pedig többnyire abból ered, hogy ezek fontosságával nincsenek tisztában (Davis, 2011). A szűrök hiánya azért is kulcskérdés, mert korunkban az információ (és az adatok) felhasználói magukra vannak hagyva, tehát önállóan kell dönteniük az információ releváns vagy irreleváns, értékes vagy értéktelen voltáról, mivel az információs kapuőrök (lektorok, szerkesztők, könyvtárosok) szerepe radikálisan lecsökkent (Badke, 2004), miközben egyre több információt és (egyelőre kisebb mértékben) adatot kéretlenül kapunk.

Az információs túlterhelésnek kiterjedt szakirodalma van (Koltay, 2017), de megjelent az adattúlterhelés is, amely az adatok hatékony menedzselése és kritikai adatműveltség (Špiranec–Kos–George, 2019) nélkül nehezen leküzdhető. Természetesen tudnunk kell, hogy az utóbbi kérdések a mesterséges intelligencia és az adatvezérelt világ viszonyának csak a tágabb kontextusát adják meg. Éppen ezért nem annyira magára a kritikai adatműveltségre, mint inkább e szókapcsolat második tagjára, vagyis a kritikai megközelítésre fogok a következőkben koncentrálni.

Az adatok és az adattudomány kritikai szemlélete

Nemcsak arra van szükség, hogy képesek legyünk az adatok megértésére, használatára és kezelésére (Qin–D'Ignazio, 2010), tehát arra, hogy uraljuk az adatkörnyezetet (Johnson, 2011), hanem kritikai szemlélettel is kell közeledjünk az adatokhoz annak érdekében, hogy lehetővé tegyük értelmezésüket, interpretálásukat és etikus használatukat (Koltay, 2015). Ez a felismerés hozta létre az információhoz és az adatokhoz kötődő különböző írástudások, műveltség- és kultúratípusok iránti igényt, ide értve a fent említett adatműveltséget is. A hozzájuk kötődő fogalomcsalád természetének, valamint a „családtagok" közötti hasonlóságoknak és különbségeknek a kifejtése (Z. Karvalics, 2012) helyett csak Lloyd (2017) véleményére térek ki, miszerint ezek alapvetően különböző társadalmi környezetben megvalósuló gyakorlatoknak tekinthetők. Az irántuk megnyilvánuló igény megjelenése

többek között annak a konvergenciának az eredménye, amelynek hatására a távközlés, a számítástechnika és az elektronikus média hálózati információs és kommunikációs technológiákként egyesültek (Livingstone–van Couvering–Thumin, 2008). Ebben a digitális környezetben ugyanis mód nyílik arra, hogy újrahasznosítsuk az információs tárgyakat (Steinerová, 2010). A technológiai konvergencia egyúttal azt is eredményezte, hogy az előbbiekben említett műveltségek (írástudások) közeledtek egymáshoz.

Ontológiai szempontból az adatok erősen kötődnek kontextusukhoz, amely nélkül értelmetlenek lennének, mivel minden adat létrejöttét kulturális és történeti környezete határozza meg. Ezt a kontextust azonban nem könnyű egyértelműen értelmezni.

Ahogy azt Neff, Tanweer, Fiore-Gartland és Osburn (2017) kifejtik, fontos figyelnünk arra, hogy különbség van a reprezentacionális és az interakcionális kontextus között. Az előbbi az adatokon kívül létező, meghatározható, leírható és kódolható környezetként jeleníthető meg, az utóbbi pedig az adatok felhasználóinak kommunikációjával és helyi gyakorlatával fonódik össze.

Az érem másik oldala, hogy az adatok kontextusai nem eleve adottak, ami megerősíti annak a megközelítésnek a helytálló voltát, mely szerint a kontextusok nem függetlenek a gyakorlattól, mivel tulajdonságok egymáshoz való viszonyaiból eredő cselekvések eredményeiként jönnek létre (Seaver, 2015).

Kontextuális jellegüknél fogva, az adatok megértése mindig interpretáción alapul, tehát egyrészt az adatok maguk is interpretációk, másrészt megértésük megköveteli az értelmezést. Az adatok fontosságának megállapítása és kommunikálása magában foglalja saját megértésünk és értelemzésünk, valamint célközönségünk fogékonyságának felbecslését. Ez azt jelenti, hogy az adatokat nem közelíthetjük meg semleges módon, ami aztán oda vezet, hogy kritikával nézzük a néha nekik tulajdonított privilegizált episztemológiai státuszt (Špiranec–Kos–George, 2019).

Az adatok episztemológiai státuszának vizsgálatakor ezért dekonstrukció útján le kell bontanunk és újra kell értelmeznünk az olyan kijelentéseket, amelyek alapján az adatoknak kétségbevonhatatlan autoritást (szakmai tekintélyt) tulajdonítunk. Nem mellesleg, a dekonstrukció, vagyis a tartalmak részekre szedhető és elemezhető voltának felhasználása a különböző műveltségek (írástudások) esetében is fontos szerepet játszik (Aczél, 2013).

Mivel az adatok nem semlegesek, természetük ideológiai (Markham, 2018). Amikor látszólag kézzelfogható igazságokat közvetítenek (például) infografikák és táblázatok formájában, akkor az adatok a világ reduktív képét nyújtják (Špiranec–Kos–George, 2019). Emellett, az adatokra alapozott kijelentések hátterében sokszor az áll, hogy valaki valamit bizonyítani akar (Tygel–Kirsch, 2016). Az ilyen adatok megbízhatósága tehát kérdéses, aminek alapján feltételezhetjük, hogy a mesterséges intelligencia számára sem nyújtanak adekvát inputot.

Amikor adatokat használunk, olvasás és feldolgozás útján értelmezzük őket. Az adatok létrehozása során viszont saját interpretációinkként konstruáljuk meg őket. Ezek a folyamatok bizonyos szempontból a befogadás és kizárás mintázataiként, többé vagy kevésbé szubjektív becsléseket jelentenek, amelyek a különböző kontextusokban az adatokkal kapcsolatos, adottnak vett normákon és szabványokon alapulnak (Neff–Tanweer–Fiore–Gartland–Osburn, 2017).

Az adatokkal kapcsolatos kritikai hozzáállásunk mozgósíthatja azt a tudástárat, amelyet akár a mesterséges intelligencia egy lehetséges, bár csak részfolyamatát érintő előképének is tekinthetünk. A tudástár rendszerezett információkészletek összessége, amelyet magunknak kell felépítenünk, majd memóriánkban tárolnunk, mivel sohasem spontán módon jön létre. Átgondolt terv szerint összeállított információelemekből épül fel, tehát jóval több, mint tények halmaza. Struktúrája segít bennünket a mintázatok felismerésében. Ezeket aztán térképként használjuk a további információk megtalálásához, és arra is jók, hogy segítségükkel előkeressük a korábban már az ismereteink közé eltárolt információkat (Potter, 2015).

A nagy adatok kapcsán fentebb megfogalmazott (és viszonylag közismert) aggodalmak mellett érdemes odafigyelnünk az algoritmizációt ért kritikákra is, amelyeknek fontos része, hogy az nagymértékben kihat, vagy kihathat a társadalomra azzal, hogy a döntéshozatalt a technológiába vetett implicit bizalom alapján automatikus rendszerekre bízza (Špiranec–Kos–George, 2019). Carrington (2018) szerint azok az identitások, amelyeket algoritmusok hoznak létre, mindenre kiterjedő módon befolyásolják életünket, mivel szakadék van aközött, akinek tartjuk magunkat és aközött, akivé az algoritmusok tesznek bennünket. Ezzel elveszik tőlünk annak a lehetőségét, hogy egy többé-kevésbé koherens önképet (narratív ént) magunk hozzunk létre. Ehelyett egyre inkább egy kizsákmányoló kapcsolat jön létre, amely egyre több személyes adatot követel meg annak elérése végett, hogy az algoritmus jobban értsen bennünket. Az algoritmusok ugyanakkor technológiai szempontból is problematikusak, viszont legitimálja használatukat a nagy adatok körüli felhajtás, amelynek kapcsán Boyd és Crawford (2012) figyelmeztetnek arra, hogy az internetről vett nagy adatállományok gyakran megbízhatatlanok, ezért ismernünk kell jellemzőiket és korlátjaikat. Ha egy adatállomány mérete milliós nagyságrendű, attól még nem feltétlenül reprezentatív. Hogy statisztikailag érvényes következtetéseket vonjunk le egy-egy adatállományból, tudnunk kell, honnan származik és melyek a gyengeségei. Tudatában kell lennünk azoknak a tényezőknek, amelyek az értelmezést torzíthatják. A nagy adatok ugyanis arra hajlamosítanak bennünket, hogy ott is összefüggéseket lássunk, ahol valójában nincsenek.

Az algoritmusokat sokan fekete dobozként írják le, tehát átláthatatlannak tekintik (Willson, 2017). Mivel működésük sok tekintetben nem átlátható, nehéz őket elemzésre használni és döntéshozatali mechanizmusaik határát meghúzni (Lloyd, 2017). Ennek sajátos példája, amikor nem tudhatjuk, hogy az adott tartalom hogyan került napvilágra pusztán azért, mert egy mesterséges neurális valamely más tartalomhoz hasonlónak találta. Ez eszünkbe juttatja a szűrőbuborékok (filter bubbles) problémakörét (Cox–Pinfield–Rutter, 2018). A szűrőbuborékok egyrészt – a felhasználói igények minél tökéletesebb kiszolgálását megcélozva – meghatározott algoritmusok révén lehetővé teszik a médiafogyasztás mintázatainak és a felhasználói preferenciáknak a megfigyelését, másrészt gyakorlatilag kizárják azokat a nézeteket, amelyek eltérnek az adott felhasználó saját véleményétől (Pariser, 2011). A fenti példa azonban némileg másról szól, mivel elsősorban nem az eredményt tekinti problematikusnak, hanem azt, hogy létrejöttének módját nem ismerhetjük megfelelő részletességgel és pontossággal.

Egyet kell tehát értenünk Thomas, Nafus és Sherman (2018) megállapításával, hogy az algoritmusok erejét az a hatalom adja, amellyel felruházzuk őket. Ezt tesszük ugyanis, ha

az egyszerű robotporszívó útvonaláról készült pillanatképről azt gondoljuk, hogy az többet jelent a ház kitakarításánál.

A mesterséges intelligencia esetében mindenesetre megjelent a döntéshozatal és az ellenőrzés kérdéseinek szabályozása iránti igény, amelyhez hasonlót az adatok esetében is ismerünk, hiszen ez a feladata az adatkormányzásnak: az adatok felhasználása előtt álló célkitűzések megvalósítása érdekében szabványosított és megismételhető folyamatokra építve átláthatóvá teszi az adatokkal kapcsolatos folyamatokat és a döntéshozatalt (DGI, 2015). A rokonság ezekben a követelményekben megvan a mesterséges intelligencia esetében is, azonban ki kell bővülniük (többek között) azzal, hogy miként tervezhetők meg és működtethetők azok a mesterséges intelligenciára épülő rendszerek, amelyek olyan humán értékekre reflektálnak, mint a tisztesség, a felelősségvállalás, az átláthatóság és az előítéletmentesség (Gasser–Almeida, 2017).

A kutatási információkkal többféle szakember foglalkozik, van köztük adatkönyvtáros (data librarian), adatkurátor (data curator) és adattudós (data scientist). Az utóbbi szakemberek elnevezése (adattudós, data scientist) már önmagában is sokak ellenérzését válthatja ki. Az elnevezés mellett és ellene is érvelhetünk, de sokkal fontosabb, hogy egyelőre azt látjuk, az adattudomány rövidtávú célokat tűz ki maga elé (Voulgaris, 2014), tehát joggal merül fel, hogy kritikával kell szemlélni.

Az adattudomány elsősorban üzleti elemzések kiterjesztése, amely számítástudományi, statisztikai és alkalmazott matematikai eszközöket használ nagy mennyiségű adat automatikus elemzésére. Pusztán diszkurzív technológiák elegye, tehát szó sincs arról, hogy felülírná a tudományosság követelményét (Buzato, 2017).

Robinson (2016) emellett arra is figyelmeztet, hogy ha kizárólag a mintázatokra és a szintaxisra figyelve, kvantitatív módszereket alkalmazunk, nehezen hozzáférhető adatsilókat építünk, ami megnehezíti az adatok megtalálását és felhasználását. Másrészt viszont úgy látja, hogy az adattudomány a rögzített információ teljes kommunikációs láncát tanulmányozza, mivel célja az, hogy a nyers adatokból hasznosítható, a döntéshozatalt segítő tudást állítson elő. Emellett azonban szükség volna arra, hogy olyan koherens és elméleti megalapozottságú keretek létrejöttéhez járuljon hozzá, amelyek egyaránt érintik az adatok műszaki-technológiai és humán dimenzióit, továbbá átfogják az emberi megértés folyamataival kapcsolatos vélekedéseket, az információtechnológiai folyamatokat és azok társadalmi hátterét, valamint az információ értelmezéséhez szükséges szociokulturális és etikai összetevőket is (Robinson–Bawden, 2017; Wang, 2018).

Az adattudomány leggyümölcsözőbb elméleti keretét az a szemlélet adná, amely azt feltételezi, hogy az információ vagy az adatok különböző társadalmi csoportokhoz fűződő, eltérő érdekeket tükröznek (Hjørland, 2019). Úgy tűnik azonban, hogy ennek az elvnek a gyakorlatba történő átültetésére még várnunk kell.

Aligha fér kétség ahhoz, hogy a mesterséges intelligencia hatékony és hibátlan működéséhez jó minőségű adatokra van szükség. Ezt erősíti meg a mesterséges intelligencia kutatását és fejlesztését támogató országos szuprastruktúra megvalósítására kidolgozott tanulmány is (Kovács–Pallinger, 2019). Az adatok minősége tehát kulcsfontosságú kérdés. Erre mutat rá az a vita is, amelynek tárgya az, hogy az adattudomány hatóköre kiterjed-e az adatminőségre.

Korábban Cao (2016) úgy látta, hogy az adattudomány figyel az adatok minőségére, míg Wang (2018) szerint főként az adatok mennyiségével törődik. Mindenesetre, ha nem figyel az adatok minőségére, akkor félő, hogy a régi mondás: "garbage in, garbage out" (ha szemetet viszünk be, az output is szemét lesz) érvényes marad a nagy adatok korában is (Wang, 2018).

Az adatok felhasználásának alapja az irántuk megnyilvánuló potenciális igény természetének és mértékének azonosítása. Amikor tehát az adatokat értékeljük, meg kell vizsgálnunk az adott adatállomány forrásának relevanciáját, az adatok kompatibilitását és az őket leíró metaadatok minőségét (Carlson et al., 2011).

Az adatminőség magába foglalja azoknak a kontextusoknak és átalakításoknak a figyelembevételét, amelyek során az adatok létrejöttek (Ramírez, 2011). A minőség megítélését ugyanakkor befolyásolják az adatok értékelőinek elfogultságai és előítéletei. Összetett, sokdimenziós jellegénél fogva az adatminőség más dimenziói is figyelmet érdemelnek. Ezek egyike a bizalom, amelynek mértéke számos szubjektív tényezőtől függ, és amely önmagában is felülírhat más szempontokat. Hasonló tényező az adatok létrehozóinak jó híre. Az adatokat ezenkívül autentikusnak kell megítélnünk, felhasználásukat vagy alkalmazásukat elfogadhatónak kell találnunk. A hitelesség ebben a kontextusban olyan kérdéseket állít a középpontba, mint az adatgyűjtés eszközeinek megbízhatósága, az elméleti alapok megfelelő volta, az adatok teljessége, pontossága és érvényessége. Annak érdekében, hogy a hitelességet meg tudjuk ítélni, az adatoknak érthetőeknek kell lenniük. Az érthetőség értékeléséhez nélkülözhetetlen, hogy az adatokat leíró dokumentáció, metaadatok vagy az adatok eredetére vonatkozó információk formájában elegendő kontextus álljon rendelkezésre, valamint az, hogy az adatok használhatók legyenek. A használhatóság megköveteli, hogy az adatok megtalálhatók és hozzáférhetők, a fájlformátumok megfelelőek legyenek, hogy az adatok minőségét megítélő egyének alkalmas eszközökkel rendelkezzenek az eléréshez, továbbá, hogy biztosítva legyen az adatok kellő mértékű integritása. Utóbbi azon a bizonyosságon alapszik, hogy az adatok teljesek és hiánytalanok, konzisztensek és helyesek mind intellektuálisan, mind technikai szempontból. Az integritást a létrehozás és a használat bármely fázisában veszélyeztethetik emberi hibák. Mivel az adatok javítása mindig költséges, a legjobb gyakorlat az, ha kezdettől fogva helyes adatokkal dolgozunk. Az integritás azt is feltételezi, hogy az adatok a bitek szintjén bizonyíthatóan azonosak legyenek egy korábbi, elfogadott és ellenőrzött állapottal (Giarlo, 2013). Az adatminőség vizsgálatában ugyanakkor a mesterséges intelligencián alapuló szűrők is használhatók, méghozzá humán szakértői bírálattal kombinálva (Kelling et al., 2015).

Összegzés

Az adatvezérelt világ és a mesterséges intelligencia kapcsolatának megértéséhez nélkülözhetetlen az előbbi természetének mélyebb megértése. Ennek érdekében ez az írás – főként a kutatási adatokra koncentrálva – először az adatok szemléletének változását mutatta be, majd szólt a nagy adatokkal kapcsolatos várakozásokról és félelmekről. A problémák között megemlítettem az információs- és adattúlterhelés kérdéskörét. Ezt követte az adatok

minőségének viszonylag részletes tárgyalása, amely felvillantotta a mesterséges intelligencia egyik lehetséges szerepét. Figyelmet szenteltem az adatok kritikai szemléletének, valamint az adattudománnyal és az algoritmizációval kapcsolatos kritikáknak. Tettem mindezt annak reményében, hogy ezekkel a gondolatokkal adalékokat tudok nyújtani a mesterséges intelligenciával kapcsolatos interdiszciplináris és kritikai gondolkodáshoz.

Irodalom

ACRL (2000): *Information Literacy Competency Standards for Higher Education.* American Library Association, Chicago, IL.

Aczél, P. (2013): Médiaműveltség. In Nagy-Király Vivien (szerk.) *Médiatudatosság az oktatásban,* Budapest, OFI, pp. 39–44.

Al-Ruithe, M.–Benkhelifa, E.–Hameed, K. (2019): A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, 23(5-6), 839–859.

Badke, W. (2004): *Research strategies: Finding your way through the information* fog. New York, NY: IUniverse

Battista, A.–Conte, J. A. (2017): *Teaching with Data: Visualization and Information as a Critical Process*. In N. Pagowsky & K. McElroy (Eds.) Critical Library Pedagogy Handbook, (pp. 147–154). Chicago, IL: Association of College &Research Libraries

Borgman, C. L: (2015): *Big Data, Little Data, No Data: Scholarship in the Networked World.* Cambridge, MA: MIT Press, 2015.

Boyd, D.–Crawford, K. (2012): *Az adatrengeteg kínos kérdései: Vitaindító egy kulturális, műszaki és tudományos jelenségről.* Információs Társadalom, 12(2), 7–23.

Buckland, M. (1991): *Information as thing, Journal of the American Society for Information Science*, 42(5), 351–360.

Buzato, M. E. K. (2017): *Critical Data Literacies: going beyond words to challenge the illusion of a literal world.* In Construções de sentido e letramento digital critico na área de Línguas/Linguagens.(pp. 119–142). Campinas: Pontes

Cao, L. (2016): *Data science: Nature and pitfalls.* IEEE Intelligent Systems, 31(5), 66–75.

Carrington, V. (2018): *The Changing Landscape of Literacies: Big Data and Algorithms.* Digital Culture and Education, 10(1), 67–76.

Carlson, J.–Fosmire, M.–Miller, C. C.–Nelson, M. S. (2011): *Determining data information literacy needs*: A study of students and research faculty. portal: Libraries and the Academy, 11(2), 629–657.

Cox, A. M.–Pinfield, S.–Rutter, S. (2018): *The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries.* Library Hi Tech. 37(3), 418–435.

Davis, N. (2011): *Information overload, reloaded. Bulletin of the American Society for Information Science and Technology*, 37(5), 45–49.

DGI (2015): *Definitions of Data Governance, Data Governance Institute,* http://www.datagovernance.com/adg_data_governance_definition/

Emmelhainz, C.–Pappas, E.–Seale, M. (2016): *Thinking through Visualizations: Critical Data Literacy Using Remittances.* In N. Pagowsky–K. McElroy, (Eds.) Critical Library Pedagogy Handbook, (pp. 179–187.) Chicago, IL: Association of College & Research Libraries

Floridi, L. (2012): *Big data and their epistemological challenge. Philosophy & Technology,* 25(4), 435–437.

Floridi (2015): *The Advisory Council to Google on the Right to be Forgotten* https://www.google.com/advisorycouncil/

Floridi, L. (2016): *Should we be afraid of AI?* Aeon Essays, https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

Floridi, L. (2017): *Charting our AI future, Project Syndicate,* https://www.project-syndicate.org/commentary/human-implications-of-artificialintelligence-by-luciano-floridi-2017-01

Frické, M. (2019): *The knowledge pyramid: the DIKW hierarchy.* Knowledge Organization, 46(1) 33–46.
Gasser, U.–Almeida, V. A. (2017): *A layered model for AI governance.* IEEE Internet Computing, 21(6), 58–62.
Giarlo, M. (2013). *Academic Libraries as Data Quality Hubs. Journal of Librarianship & Scholarly Communication,* 1(3), 1–11.
Hjørland, B. (2019): *Data (with Big Data and Database Semantics).* Knowledge Organization 45(8): 685–708.
Jones, Q.–Ravid, G.–Rafaeli, S. (2004): *Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration.* Information Systems Research, 15(2), 194–210.
Kelling, S.–Fink, D.–La Sorte, F. A.–Johnston, A.–Bruns, N. E.–Hochachka, W. M. (2015): *Taking a 'Big Data' approach to data quality in a citizen science project.* Ambio, 44(4), 601–611.
Koltay, T. (2015): *Data literacy: in search of a name and identity. Journal of Documentation,* 71(2), 401–415.
Koltay, T. (2017): *Egy „örökzöld téma", az információs túlterhelés.* Információs Társadalom, 17(3), 39–54.
Kovács, L.–Pallinger, P. (2019): *Mesterséges intelligencia kutatás-fejlesztés támogató országos szuprastruktúra létesítése és adatkezelési funkciói.* Budapest: SZTAKI, Számítástechnikai és Automatizálási Kutató Intézet
Lloyd, A. (2017): *Information literacy and literacies of information: A mid-range theory and model. Journal of Information Literacy,* 11(1), 91–105.
Livingstone, S.–van Couvering, E. J.–Thumin, N. (2008): *Converging traditions of research on media and information literacies*: Disciplinary and methodological issues. In J. Coiro, M. Knobel, C. Lankshear, & D. J. Leu (Eds.), (pp. 103–132.) Handbook of Research on New Literacies. New York, NY: Routledge
Makani, J. (2015): *Knowledge management, research data management, and university scholarship: Towards an integrated institutional research data management support-system framework.* VINE, 45(3), 344–359.
Morville, P (2005): *Ambient findability.* Sebastopol, CA: O'Reilly
Neff, G.–Tanweer, A.–Fiore-Gartland, B.–Osburn, L. (2017): *Critique and contribute: a practice-based framework for improving critical data studies and data science.* Big Data, 5(2), 85–97.
Pariser, E. (2011): *The Filter Bubble. What the Internet is hiding from you. N*ew York, NY.: The Penguin Press
Peirce, Ch. S. (1960): *Collected Papers of Charles Sanders Peirce, Vol. 2.* Cambridge: Harvard University Press
Pijpers, G. (2010): *Information Overload: A System for Better Managing Everyday Data.* Hoboken, N. J.: Wiley
Potter, W. J. (2015): *Médiaműveltség*. Budapest, Wolters Kluwer
Pryor, G (2012): *Managing research data.* London: Facet
Qin, J.–D'Ignazio, J. (2010): *Lessons learned from a two-year experience in science data literacy education.* International Association of Scientific and Technological University Libraries, 31st Annual Conference. Paper 5. http://docs.lib.purdue.edu/iatul2010/conf/day2/5
Ramírez, M. L. (2011): *Opinion: Whose role is it anyway? A library practitioner's appraisal of the digital data deluge.* Bulletin of the American Society for Information Science and Technology, 37(5), 21–23.
Robinson, L. (2016): *Between the deluge and the dark age: Perspectives on data curation.* Alexandria, 26(2), 73–76.
Robinson, L.–Bawden, D. (2017): *"The story of data" A socio-technical approach to education for the data librarian role in the CityLIS library school at City, University of London.* Library Management, 38(6/7), 312–322.
Rowley, J. (2007): *The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science*, 33(2), 163–180.

Seaver, N. (2015): *The nice thing about context is that everyone has it*. Media, Culture & Society, 37(7), 1101–1109.

Špiranec, S.–Kos, D.–George, M. (2019): S*earching for critical dimensions in data literacy.* Information Research, 24(4), paper colis1922. http://InformationR.net/ir/24-4/colis/colis1922.html

Steinerová, J. (2010): *Ecological dimensions of information literacy.* Information Research, 15(1), colis719

Thomas, S. L.–Nafus, D.–Sherman, J. (2018): *Algorithms as fetish: Faith and possibility in algorithmic work. Big Data & Society,* 5(1), 2053951717751552.

Tygel, A. F.–Kirsch, R. (2016): *Contributions of Paulo Freire to a critical data literacy: a popular education approach.* The Journal of Community Informatics, 12 (3), 108–121.

Voulgaris, Z. (2014): *Data Scientist: The Definitive Guide to Becoming a Data Scientist.* Basking Ridge, NJ: Technics Publications

Wang, L. (2018): *Twinning data science with information science in schools of library and information science.* Journal of Documentation, 74(6), 1243–1257.

Willson, M. (2017): *Algorithms (and the) everyday. Information, Communication & Society,* 20(1), 137–150.

Yu, L. (2015): *Back to the fundamentals again. Journal of Documentation,* 71(4), 795–816.

Z. Karvalics, L. (2012): *Információs kultúra, információs műveltség – egy fogalomcsalád értelme, terjedelme, tipológiája és története.* Információs Társadalom, 12(1), 7–43.