A természetes intelligencia manifesztációjának filozófiai kérdései*

Prisznyák Alexandra

Héder Mihály:
Mesterséges intelligencia – Filozófiai kérdések, gyakorlati válaszok
Gondolat Kiadó, Budapest, 2020, 166 o.
ISBN: 9789635560509

Héder Mihály habilitált egyetemi docensként és technikafilozófusként a természetes nyelvek feldolgozásával, szemantikus annotációs technológiák és rendszerek tervezésével, továbbá a mesterséges intelligencia filozófiájával foglalkozik.

A könyv gondolatmenete a mesterséges intelligenciával (AI) összefüggő filozófiai kérdések köré szerveződik. A modern „best seller" könyvek témájához hasonlóan, részletesen kitér az AI fejlődési korszakait érintő kritikákra, filozófiai, etikai kérdésekre, és kapcsolódóan tárgyalja a mesterséges intelligencia potenciális jövőbeni lehetőségeit. A könyv sikerét elsősorban a téma aktualitása, nevezetesen az AI, robotok piaci implementációja szolgáltatja, filozófiai megközelítésben.

A humán munkavállalók mind inkább emberibb munkavégzését támogató AI történelmi fejlődésének kezdeteként általában az 1956-ban megrendezett darthmouh-i Artificial Intelligence konferenciát jelölik meg. Szűkebb körben ismert, hogy a gépi intelligenciával kapcsolatos munka Nagy-Britanniában már közel egy évtizeddel korábban elkezdődött, az Enigma elleni világháborús erőfeszítéseknek köszönhetően. *Alan Turing* 1950-ben a Computing Machinery and Intelligence c. publikációjában ráirányította a figyelmet az intelligens gépekben rejlő potenciálra, amely azóta is vita tárgyát képezi. A Turing-teszt néven emlegetett imitációs játék a gépek „gondolkodásának[1]" mérését célozza. Az AI-teszten elért siker alátámasztja az AI performatív

* A jelen kiadványban megjelenő írások a szerzők nézeteit tartalmazzák, ami nem feltétlenül egyezik a Magyar Nemzeti Bank hivatalos álláspontjával.

Prisznyák Alexandra: Nemzetközi Bankárképző Központ Zrt., szenior tanácsadó és mesterséges intelligencia & CBDC Program Manager; Pécsi Tudományegyetem, doktorjelölt. E-mail: aprisznyak@bankarkepzo.hu

[1] A gondolkodás fogalmával kapcsolatosan Turing úgy tartja, hogy az rosszul definiált, túl általános, így helyette a Turing-tesztet ajánlja. Turing azonban sohasem állítja, hogy a gépek sikeres működése egyenlő a gondolkodással. A teszt egy háromszereplős imitációs játék, amelyben a bírón kívül – aki nem hallja és nem látja a másik két szereplőt – két ágens (egy ember és egy digitális számítógép) vesz részt. A játék célja a bíró félrevezetése, aminek során mindkét fél hazudhat. A véletlen találatok elkerülése érdekében több bírót és kritériumot alkalmaznak. Célja, hogy a gép képes legyen a bírók több mint 70 százalékát megtéveszteni és elhitetni velük, hogy ez ember.

kudarcát jósoló pesszimisták tévedését. Ugyanakkor egyben felveti azt az alapvető problémát, amely szerint az ember nem képes saját fajának felismerésére.

Az első fejezet számos AI-kritikus munkásságát és a kapcsolódó kritikai észrevételeket elemzi. *Joseph Weizenbaum* felfogásában a számítógép nem más, mint a megtestesült szisztematikus matematika, aminek az emberi intelligenciához nincsen semmi köze, csupán a programozók által megírt algoritmikus utasításokat követi. Ezt bizonyítandó, 1964-ben az ember intellektuális képességét és a gépek korlátozott képességeit demonstrálni hivatott számítógépes program írásába kezdett. Az ELIZA névre keresztelt chatbot kifejlesztésével Weizenbaum célja az volt, hogy a beszélgetőtársról (chatbot) kezdetben kialakított pozitív benyomás – a társalgás előrehaladtával (amikor a gép korlátaira a váratlan szituációban fény derül) – egyre inkább az illúzió/emberrel történő kommunikáció benyomását keltse. Szándékaival ellentétben a piac lelkes fogadtatása sokkolta Weizenbaumot. A szakmai közeg és a piac is kulcsmomentumként értékelte ELIZA létrejöttét, amely a fejlődés tekintetében számos ajtót kinyithat, többek között a természetes nyelvek számítógépes feldolgozásának területén. Weizenbaum karrierjének hátralévő részében a gépeknek való alárendelődés veszélyére kívánta felhívni a figyelmet, s kritizálta az AI-kutatásokat. A szerző Weizenbaum példáján keresztül illusztrálja a kor AI-ba vetett hitét, a fellendülés korszakát, amely később elvezetett az első és második AI-télhez[2], végezetül pedig napjaink dinamikusan fejlődő digitális érájához és az AI térnyeréséhez.

A kritikai észrevételek típusainak rendszerezéséhez a szerző szükségképpen kitér az AI-kritikák három fontos területére: a performatív és a fenomenológiai kudarcra, illetve az AI fejlődéséhez kapcsolódó disztópiára. A performatív kudarc a mesterséges intelligencia megoldások működésképtelenségére irányuló észrevétel, amely az AI-telek alatt számos esetben jelentkezett, megtörve a fejlődés lendületét. A kritikák alapját az az álláspont szolgáltatja, amely szerint a gépek nem alkalmasak az intelligens viselkedést igénylő feladatok elvégzésére. A fenomenológiai kudarc alapjait az AI – korábbiakban is említett – „imitáló jellege" szolgáltatja. Az emberi elme működési mechanizmusától (biokémiai elv) eltérő gépi működés (bináris-kivételt a kvantumszámítógépek jelentenek) eredményezi azt a kritikai típust, amely szerint a tudat élménye (így például az érzelmek átélése) nem áll elő a gépek számára (lásd például Searl kínai szoba gondolatkísérlete). Turing a fenomenológiai kudarccal összefüggésben tévesnek tartja azt a vizsgálódási kiindulópontot, amely kizárólag azon alapszik, hogy a gépek működése nem emberszerű. Nem állítja tehát, hogy a humán működéshez hasonlóak lennének és ugyanazt az élményt élik át, mint az intelligens emberek. Kizárólag arra irányítja a vizsgálódás fókuszát, hogy a gondolkodás nem kellőképpen definiált egy ilyen következtetés levonásához. A fogalom

[2] Az AI tél egy olyan időszakot jelöl, amikor a mesterséges intelligencia fejlesztési eredmények stagnálnak, elmaradnak a befektetői elvárásoktól, amelynek eredményeképpen a fejlesztések leállításra kerülnek és a közvélemény eltávolodik az AI technológia iránti lelkesedéstől.

tartalmi vonatkozása ugyanakkor az intelligens gépek megjelenésével szükségképpen igényli a definíció újragondolását. Így a fenomenológiai megközelítés önkényes kritikai észrevételként értelmezendő. A szerző egy roppant érdekes tényre hívja fel a figyelmet e tekintetben: „[…] ha a fenomenológiai siker lehetetlen, akkor a tudatfeltöltés egyenlő a tudat halálával".

Az AI-kritikusok egyik legaktívabb képviselőjeként tevékenykedő *Hubert Dreyfus* filozófus a mesterséges intelligenciát a ráfordított erőforrások elvesztegetésének vélte. Meglátása szerint a gépek csak előre beprogramozott szabályok mentén képesek operálni. Így a szabály alapú, utasítások sorát végrehajtó gép sosem lesz képes humán tulajdonságok produkálására. Álláspontját az AI-telek alatt megvalósult kutatási projektek kudarcaival támasztja alá. Az ennek ellenére szakadatlanul tartó kutatói kedvet négy előfeltevéssel magyarázza: (1) biológiai (az emberi agy információfeldolgozása diszkrét egységekben valósul meg, így léteznie kell olyan biológiai elemnek, amely a digitális technika elemeire hasonlít); (2) ontológiai (információ-hozzáférhetőség); (3) pszichológiai (az elme formális szabályokon alapuló információfeldolgozása), (4) ismeretelméleti előfeltevés (minden tudás formalizálható, Boole-függvények). Dreyfus véleménye mellett még akkor is kitartott, amikor a gép (IBM Watson) első ízben aratott győzelmet az ember felett a Jeopardy! vetélkedőben 2011-ben. Dreyfus négy előfeltevése azon a – téves – megállapításon alapul, amely szerint az emberi elme és a számítógép egyaránt általános szimbólumfeldolgozó gép, s álláspontját erős érvekkel támasztja alá az emberi tudás formalizálhatóságának megkérdőjelezésével. Mindazonáltal az általános szimbólumfeldolgozó gép tulajdonságai a Turing gépre jellemzőek (ismeretelméleti megértés és tervezés), így az azzal történő azonosítása éppúgy hiba, mint az embert azonosítani vele. Turing azonban nem azonosítja saját modelljében a két dolgot. Dreyfus érvelésében kitér a digitális számítógépek (fizikai megjelenés hiányában jelentkező) észleléssel és cselekvéssel összefüggő hiányosságaira. Az érvelésének hibás alapjaira a világbajnok *Garry Kaszparovot* legyőző Deep Blue számítógép az „élő" bizonyíték.

A természetes intelligencia manifesztációjának mibenlételeként (imitáció) jelentkező problémakör a gépek tudatosságának képességét vizsgálja. A belső élmény megélése, abszorpciója nélkül (tudat) működő gépek alapjaiban véve különböznek az intelligens humán viselkedéstől. *John Searle* a gépek megértésének képességét tagadja, amikor megalkotja a kínai szoba gondolatkísérletet. Feltételezése alapján elképzelhetetlen a gépek megértési képessége a performatív és fenomenológiai siker egyidejű teljesülésének hiányában. A megértés élményével kapcsolatos ellenérvek a gondolatkísérlet számos aspektusát célozták, így az izolációt, az ágens határait (az emberi agy kiragadott szelete vajon képes-e a megértésre), extrapoláció, tanulás, memória, egyebek, amelyek a megértéssel hasonló intellektuális élmény kialakulását eredményezhetik. Ellenérvként a szakértők az intelligens robotok esetében a szenzorokkal, érzékelőkkel történő felruházást hozzák fel. Searle

gondolatkísérletéhez csatlakozva a könyv második fejezetébe a szerző részletezi a *Daniel C. Denett* „intuíciópumpáinak" elve alapján megfogalmazott legfőbb kritikáját: az időtényezőt.

A kritikák egy másik dimenzióját képviseli az AI elterjedésével kapcsolatos munkaerő-felszabadítás mértékétől való félelem. A szerző az AI-eszkaláció episztemikus korlátaival kapcsolatosan hangsúlyozza, hogy az AI egy ponton túl a tudás bővülésének korlátjába ütközik. Így egy hirtelen bekövetkező eszkaláció igencsak valószínűtlen. Az AI fokozatos bevonása a munkafeladatokba több szempontból is jelentősen meghatározott. Az optimalizálás, a termelékenység növekedése, a munkahelyteremtés/rombolás (feladatok ismételt felosztása) következtében létrejött súrlódás hozzájárulhat a termékek/szolgáltatások árának csökkenéséhez, és ezzel párhuzamosan a hozzáférhetőségük javításához – a fogyasztásból korábban kirekesztett társadalmi csoportok számára is. Ugyanakkor az egyre szélesebb körű alkalmazása munkaerőpiaci polarizációhoz vezet. Direkt hatása révén a jól definiálható, szabályokkal leírható (monoton, rutin-) feladatok kiváltásra kerülnek, míg indirekt hatása révén új jellegű feladatok jönnek létre. Következésképpen elkerülhetetlen a humán-robot interakció egyre intenzívebbé válása, esetenként még olyan szakterületeken is, ahol bár a feladat jellege az automatizálást eredményezné, mégis a szakértő bevonása és a munkafolyamat felülvizsgálata igényli a humán oldali résztvevő jelenlétét az interakciós folyamatban. A munkafolyamat újfajta felosztásának szűk keresztmetszeteként jelentkezik: (1) az észlelés és manipuláció képessége, (2) a kreatív intelligencia, (3) a társas intelligencia. A mesterséges intelligencia fejlődési ütemének állandósága nagymértékben meghatározza, hogy a jelen korban utópisztikusnak feltüntetett szuper mesterséges intelligencia elérése milyen messze van időben. Ezzel párhuzamosan olyan alapvető kérdéskörök is jelentkeznek, amelyek a humán mellett megjelenő mesterséges intelligencia jogaira irányulnak. Így már értelmet nyerhet az intelligens AI önvédelemre, lekapcsolás elleni védelemre irányuló felhatalmazása.

Az első fejezetben megfogalmazott kritikai észrevételeket a szerző a második fejezetben taglalt performatív sikerekre hozott példákkal és alkalmazási területekkel szemlélteti. Így többek között részletesen ismerteti a SHRLDU (program), Shakey robot, MYCIN (program), Herbert (robot), sakkprogramozás területén elért sikereket. A szerző kitér továbbá a neurális hálózatoknak a mesterséges intelligencia virágzásában betöltött kiemelkedő szerepére, majd a gépek természetének vizsgálásán keresztül a számítások természetét elemzi. Az evolúciós eljárások a számítások napjainkban egyik legfejlettebb típusaként elvezetnek az önmagát is dinamikusan változtatni, módosítani képes programokhoz. Példaként az Evolúciós Game of Life és AARON (a festőrobot) kerül részletezésre, amelyek igazolják, hogy amennyiben a számítás kimenetét nem fizikai jelleggel is bíró eredménytermékként fogadjuk el, úgy az imént említett példákat nem tudjuk értékelni. Egyes esetekben (például

sakk) a kimenet értelmezése szimbolikusan reprezentált adatként is megvalósulhat. A formalizáció kérdéskörét vizsgálva a szerző megállapítja, hogy számítással nem kizárólag formalizált problémák megoldására nyílik lehetőség.

Az utolsó fejezet az etika, az amorális ágens, illetve az AI jelenét és jövőjét tárgyalja. A mesterséges intelligencia morális képességének elvetése automatikusan együtt jár a felelősségviselés képességének elvetésével is. Következésképpen az AI-t előállító, üzemeltető, birtokló cégeknek foglalkozniuk szükséges az etikai szabályozás kérdéseivel és a felelősség újraosztásának következtében viselendő materiális (anyagi kártérítés) /immateriális következményekkel (reputáció). A szerző álláspontja szerint az etikai felelősségvállalás tekintetében számos piaci szereplő, így a szabályozó, a gyártó, karbantartó/üzemeltető, sőt a fogyasztó együttes felelősségvállalása is szükséges. A fenomenológiai és performatív kritikák tekintetében megvalósuló konszenzus fontos mérföldkő lehet a további előrehaladásban.

Héder Mihály könyve a mesterséges intelligencia és kapcsolódó technológiák üzleti szempontból történő megközelítéséhez releváns kiindulópontot szolgáltat. A könyv a közgazdász és jogász közösség figyelmébe ajánlható, hiszen tartalma számos már meglévő és új kutatási területen, így az AI termelékenységre és munkaerőpiacra gyakorolt hatása, az AI-rendszer tervezésének üzleti és etikai szempontjai, az AI szervezeti kultúrába integrálásának kihívásai (változásmenedzsment), a vizionárius menedzsment szerepe, az AI jogi és technológiai kérdései, az AI által eredményezett hatékonyság mérési lehetőségei, az AI vállalati értékre gyakorolt hatása vonatkozásában és egyéb számításba jövő területen kihívás elé állítja a kutatókat.