Emberarcú mesterséges intelligencia

Kampis György

Eötvös Loránd Tudományegyetem Természettudományi Kar, Etológia Tanszék, Budapest, Magyarország
E-mail: kampis.george@gmail.com

Beérkezett: 2022. július 11.; elfogadva: 2022. augusztus 29.

Összefoglalás

A jelen írás alapja a témában tartott előadásom. Először általános kérdésekkel foglalkozom, majd a tervezett „EU AI Act"-ről lesz szó, utána egy VW projektet ismertetek röviden, majd a „megmagyarázható MI"-ről fogok beszélni, aztán egy saját, hazai kezdeményezésről, az Alfi projektről teszek említést. Végezetül egy kitekintés zárja le az írást.

Kulcsszavak: emberarcú MI, EU AI Act, AI Fora projekt, megmagyarázható MI, Alfi projekt

Humane AI

George Kampis

Eötvös Loránd University Faculty of Science, Dept. of Ethology
Budapest, Hungary

Summary

This writing is based on a lecture on the topic. In my other (German) affiliation I am manager of a large-scale EU project called "HumanE AI Net" (funded with 12m Euro) comprising 53 leading EU institutions, including large universities (UCL London, LMU Munich, Sorbonne, Sussex or ELTE), networks of research institutes (Fraunhofer, Max Planck Gesellschaft, INRIA, CNR Italy), large international companies (ING Bank, SAP, Philips, Airbus), etc. In the writing I discuss general issues related to Humane AI, the planned EU AI Act, social credit systems, explainable AI, and the Alphie project, respectively.

In April 2021, the European Commission proposed a regulation on artificial intelligence, known as the *AI Act*. The regulation aims at human-faced AI in a European dimension. Although it is still only a draft, the stakes are high. The planned law has, however, faults (I maintain here), to be corrected before the text passes as law.

Another subject to discuss is the study – and prohibition (at least in Europe) – of *social credit systems*. The original "Social Credit System" is a national credit rating and blacklist developed by the Government of the People's Republic of China. Proponents of the system claim that it helps regulate social behaviour, improves citizens' 'trustworthiness' (which includes paying taxes and bills on time) and promotes the spread of traditional moral values. Critics of the system, however, argue that it goes far beyond the rule of law and violates the legitimate rights of people – in particular, the right to reputation, privacy and personal dignity – and that it can be a tool for extensive government surveillance and suppression of dissent.

"Explainable AI" (XAI) has become a hot topic in recent years. AI applications are mostly "opaque": this is especially true for learning systems and by definition for neural networks (NN). The current fashion, "deep learning", usually means the application of a particularly opaque NN anyway. It is natural not to know what the system is doing and why. So, let's change that! With this tenet, XAI was born. I review some solutions to the problem.

In the writing I also mention an application, *Alphie*, the first version of which was done in the OTKA project "Good Mobile" and is now supported by the MI National Laboratory. Alphie is a science-based playful application for children that helps them to use digital tools more consciously and within limits, while developing a variety of skills. It performs the functions of a 'grandmother' who shows emotions towards the child: can be e.g. angry, loving, etc. The application makes the corresponding sounds (!) and facilitates real social interactions (e.g. sends the child to play football (!).

Keywords: humane AI, EU AI Act, AI Fora project, explainable AI, Alfi project

Hogy a címbeli kifejezés mit takar, azt igazából senki sem tudja. Persze, ha meghatározni nem is tudjuk,[1] azért jelentősen pontosíthatjuk.

Először általános kérdésekkel foglalkozom (ha nem is a meghatározás, de a jellemzés szándékával), majd a tervezett „EU AI Act"-ről lesz szó, utána egy VW projektet ismertetek röviden, majd a „megmagyarázható MI"-ről fogok beszélni, aztán egy saját, hazai kezdeményezésről, az Alfi projektről teszek említést. Végezetül egy kitekintés zárja le a jelen írást.

Mi a „Humane AI"?

A „Humane AI" szó szerinti fordításban az emberarcú MI-t jelenti *(Cataleta 2021)*. A fentiek szerint nem tervezem ezt pontosan definiálni. Az azonban biztos, hogy most „itt legel a nyáj": lépten-nyomon találkozhatunk ezzel a kifejezéssel. Akkor viszont én mi jogon teszem (előadásom, majd most az) írásom címévé? Nos, második állásomban a DFKI („Német MI kutatóközpont", értelmileg kb. a német SZTAKI) tudományos főmunkatársa vagyok. Részlegünk koordinál többet között egy ugyanezzel a címmel futó nagyobb EU-projektet,[2] amely 12 millió euró támogatással 53 vezető európai MI laboratóriumot fog össze, közte nagy egyetemeket (UCL London, LMU München, Sorbonne, Sussex vagy az ELTE), kutatóintézetek hálózatait (Fraunhofer, Max Planck Gesellschaft, Inria, CNR Olaszország), nagy nemzetközi cégeket (ING Bank, SAP, Philips, Airbus) stb. Persze 3 (vagy 4) évre és 53 részre osztva nem sok jut belőle egy-egy résztvevőnek, tehát a részvétel inkább csak a presztízs miatt éri meg.

Az alapvető kihívás olyan robusztus, megbízható mesterségesintelligencia-rendszerek kifejlesztése, amelyek képesek az ember „megértésére", az összetett valós környezethez való alkalmazkodásra és az összetett társadalmi környezetben való megfelelő interakcióra. A cél olyan mesterségesintelligencia-rendszerek létrehozása, amelyek az emberi képességeket fejlesztik, és az emberi autonómia és önrendelkezés tiszteletben tartása mellett az egyének és a társadalom egészének jogait erősítik. Az alábbiakban néhány olyan kutatási területet sorolunk fel, amelyen jelenleg jelentős hiányosságok vannak, és amelyekkel a projekt során foglalkozunk:

– A felelős MI jogi és etikai alapjai,
– „Ember a hurokban": gépi tanulás, érvelés és tervezés emberi részvétellel,
– Multimodális érzékelés és modellezés,
– Az ember és az MI együttműködése és interakciója,
– Társadalmi tudatosság.

[1] Karl Popper tudományfilozófus szerint a tudomány nem kedveli a meghatározásokat. De ezt most nem azért nem tudjuk… Jómagam amúgy szintén tudományfilozófus vagyok, azonban (főleg régebben és most legújabban is) még inkább tudományos kutatónak tartom magam – a szó alternatívája, a „tudós" magyarul rettentő rosszul hangzik, borzasztó nagyképűségnek, holott a használója nem biztos, hogy erre gondolt.

[2] A projekt menedzsere vagyok.

Lássunk néhány példát!

Az emberarcú MI lehetővé kell tegye a munkavállaló számára, hogy új karrierbe kezdjen. Ez az elképzelés az intelligens asszisztensek olyan jelenlegi elképzelésére épül, amely a felhasználó cselekedeteit és a környezet helyzeteit észlelve figyelmeztetéseket, javaslatokat és támogató információkat kínál a felhasználónak. Az ilyen rendszerek segítik az embereket abban, hogy munkájukat hatékonyabban és kevesebb hibával végezzék. Az emberi képességek bővítésének fogalma azonban messze túlmutat azon, hogy az embereket pusztán abban segítse, hogy jobban végezzék munkájukat. Ehelyett az a cél, hogy az ember olyan tevékenységek elvégzésére is képessé váljon, amelyekre egyébként egyáltalán nem lenne képes. Analógia gyanánt tekintsük egy egyszerű kotrógép és egy exoskeleton (külső váz) összehasonlítását. Az emberek könnyen áshatnak gödröket kotrógép nélkül is, de gyorsabban és mélyebbre áshatnak, ha történetesen van kéznél egy. Az exoskeleton viszont azt teszi lehetővé, hogy egy lebénult ember is járni tudjon, amire egyébként nyilvánvalóan egyáltalán nem lenne képes. Továbbá, míg a kotrógép olyan eszköz, amelyet az embernek tudatosan és kifejezetten kell működtetnie, addig egy exoskeleton szinergikusan felerősíti az emberi cselekvéseket, és implicit módon támogatja a felhasználót. Jelenleg az „ipar 4.0" (Industrie 4.0) részeként különböző mobil és viselhető rendszereket alkalmaznak, többek között a futószalagos munka támogatására. Ezek lehetővé teszik a dolgozók számára, hogy gyorsabban és jobban végezzék munkájukat, de nem változtatják meg alapvetően annak a jellegét, amit ezek az emberek ténylegesen végeznek. A projektben azonban egy olyan „intellektuserősítő" MI vízióját követjük, amely olyan szintre fejleszti az ember kognitív képességeit, hogy az, aki elveszítette korábbi munkáját, gyorsan el tudjon helyezkedni egy magasabb képzettséget igénylő, új munkakörben. Így például egy korábbi futószalagmunkásból olyan műszaki támogató személy válhat, aki a mesterséges intelligenciára támaszkodva leküzdi azt a korlátozást, amely egyébként megakadályozná őt abban, hogy ezt a munkát végezze.

Az emberarcú MI lehetővé teheti, hogy egy robotot emberi társai a „hónap csapattársának" válasszanak. Bár a mai robotokat rendszeresen bevetik veszélyes környezetben, jellemzően egyszerű feladatokra használják őket, kevés önállósággal. Így egy felderítő robotot küldhetnek képeket készíteni egy veszélyes zónából, jellemzően erős emberi távfelügyelet (ha nem is teljes távvezérlés) mellett. Ezzel szemben mi olyan rendszereket képzelünk el, amelyek a beavatkozó csapat teljes értékű tagjaivá válhatnak. Ez egyrészt azt jelenti, hogy a robotok nagy fokú önállósággal képesek lehetnek cselekedni. A csapat tagjaként a robotot utasíthatják például, hogy „menjen, és gondoskodjon a jobb oldali területről", ami azt jelenti, hogy önállóan kell felderítenie egy ismeretlen, dinamikus, strukturálatlan környezetet, meg kell tennie

minden szükséges intézkedést például a tűz megfékezéséhez, majd segítenie kell az áldozatoknak (ez magában foglalja az etikai kérdések kezelését, amelyek azzal kapcsolatosak, hogy esetleg prioritást kell felállítani, hogy kivel kell először foglalkozni vagy mit kell tenni). Másrészt azt feltételezi, hogy a csapat társadalmi struktúrájába és csoportdinamikájába be kell illeszkednie és cselekednie kell. Ez magában foglalja az érzelmeket, feszültséget és feszültséget jelző finom jelek „olvasását", mind az egyéni, mind a kollektív szinten, majd a legmegfelelőbb módon kell reagálni. Az ilyen rendszerek korántsem korlátozódnak a tűzoltókhoz hasonló vészhelyzeti erőkre; analóg módon alkalmazhatók ezek például építőipari vagy orvosi csapatoknál, hajószemélyzetnél, vagy akár tudósokból vagy technikusokból álló csapatoknál is.

Az emberarcú MI az embereket sokszínű információnak kell kitegye, ami a vitás kérdésekről a jobban megalapozott véleményeknek kedvez. Az emberek hajlamosak ugyanis a meglévő véleményükkel és meggyőződésükkel összhangban lévő információkat keresni, ez az úgynevezett „megerősítési torzítás" (confirmation bias). Ezt használják ki az online platformok az információkereséshez, valamint a közösségi hálózatok és a média, amelyek ajánló algoritmusokat alkalmaznak a felhasználók figyelmének felkeltésére. Mellékhatásként a jelenlegi platformok felerősítik és megerősítik az egyéni elfogultságot, ami a vélemények szélsőséges polarizációját és társadalmi szintű „buborékokat" eredményez, ami drámaian negatív következményekkel jár a demokrácia táplálásához szükséges plurális nyilvános vitára nézve. Emellett az információkhoz való hozzáférést gyakran rosszindulatúan torzítják a kereskedelmi vagy politikai indíttatású befolyásolók is. Az emberközpontú MI egyértelmű társadalmi dimenzióval kell rendelkezzen, ami segíthet a hírekhez és információkhoz való hozzáférés újszerű platformjainak és mechanizmusainak kialakításában, amelyek a beépített megerősítési torzítások ellensúlyozására összpontosítanak, és átlátható, intelligens módon igyekeznek az embereket a különböző új véleményekkel megismertetni. Olyan mechanizmusokat képzelünk el, amelyek segítik az egyének és közösségek tájékozódását az ellentmondásos kérdésekben, azáltal hogy több nézőpontot kínálnak, összekapcsolják az ellentétes nézeteket és az egymásnak ellentmondó érveket, és elősegítik a kritikus gondolkodást. Például egy „chatbot" (beszélgető robot) egy csoportos beszélgetésben közölhet olyan információkat, amelyek korábban nem álltak a csoport rendelkezésére, vagy pedig a fontos információk addig figyelmen kívül hagyott forrásait javasolhatja figyelembe venni. A megmagyarázható mesterséges intelligencián alapuló ember-gép interakciós modellek fejlődése újszerű kognitív kompromisszumokat hozhat létre a „confirmation bias", valamint az újdonság és a sokszínűség iránti kíváncsiság között, lehetővé téve a fenntarthatóbb és emberibb információs ökoszisztémák kialakulását.

Az EU MI törvény („AI Act")

2021 áprilisában az Európai Bizottság javaslatot tett a mesterséges intelligenciáról szóló, MI törvény néven ismert rendeletre.[3] A rendelet az emberarcú MI-t célozza meg, európai dimenzióban *(Veale–Borgesius 2021)*.

Noha még csupán egy tervezetről van szó, de a tétje mégis hatalmas. Amikor a törvény életbe lép majd, megszűnik minden rugalmasság, és az azt alkalmazó jogászok többé nem fognak (nem is lehet nekik) tartalmi kérdésekről vitatkozni, sőt várhatóan az alkalmazók nem is lesznek a téma szakértői – a jognak, a törvények alkalmazásának azonban igen. (Ebben is a GDPR-re hasonlít a dolog, ez szintén jelentős viták után került elfogadásra, de elfogadásra került, és most az alkalmazása – és a *jogi* értelmezése – van napirenden.)

Az MI törvény amúgy a szankciók között pénzbírságot ír elő, ennek maximális összege (két kategóriában) 20, illetve 30 millió euró, amit a törvényt be nem tartó cégeknek (az MI „termékek" kibocsátóinak) kell majd fizetni. Előbbre ugorva: „termék" minden, amit nyilvánosságra hoznak, ennek tehát nem kritériuma, hogy pénzt kérjenek érte. (Termék tehát lehet például egy GitHub-ról vagy máshonnan letölthető ingyenes forráskód is…) Bár a törvényalkotó a bírság maximumánál nyilvánvalóan a nagy nemzetközi gigacégekre gondolt, amelyek a termékeiket jellemzően (sok) pénzért terjesztik, elvben előállhat olyan helyzet is, hogy egy céget elmarasztalnak, amely a terméken egy fillért nem keresett, sőt, még a fejlesztési költségei sem térültek meg.

Nos, a fent említett projektünk egy jogi oktató előadást (tutorialt) tartott az MI törvényről, 2021 júniusában (!).[4] Az oktatást egy szakképzett MI jogász (Mireille Hildebrandt, VU Bruxelles) végezte. A 12 részes kurzus áttekintette a meghatározásokat és tennivalókat.

Szükséges amúgy a véleményem szerint műszakilag beleszólni a törvény vitájába. A törvényjavaslat az MI fogalmát ugyanis túl szűken határozza meg. Eszerint az MI egy tanuló rendszer. (Nyilvánvaló azonban a hozzáértők számára, hogy az MI valójában nemcsak egy tanuló rendszer lehet.) Másfelől, a törvényjavaslat az MI-t „szoftver megoldásként" definiálja. Világos azonban az is, hogy minden szoftver megoldáshoz tartozik egy ekvivalens hardver megoldás (amely tehát ugyanazt csinálja, csak nem szoftverben). A törvényt megkerülni szándékozóknak ezután nem volna más dolga, mint a maguk MI-termékét nem tanuló rendszer formájában vagy nem szoftveres megoldásként létrehozni, és az máris kikerült az MI törvény (jelenlegi formájának) hatása alól. (A hardveres megvalósítások persze néha a gyakorlatban nem kivitelezhetők: a térképes alkalmazásokról pl. meg-

[3] A szöveg jelenleg érvényes verziója megtalálható itt: https://artificialintelligenceact.eu.

[4] Ennek az anyagai ugyan „nem nyilvánosak", de nálam ingyenesen megkaphatók: george.kampis@dfki.de.

szoktuk, hogy szoftverek, és hogy nem kell emiatt egy külön célhardvert magunkkal cipeljünk…)

A tervezett törvény ugyanakkor túl tág is. Nem életszerű az MI törvény tervezetének azon kitétele például, hogy egy termék létrehozója egyetemleges felelősséggel rendelkezzen a termék *minden* használata iránt. Nyilvánvalóan itt a nem rendeltetésszerű használat helyzete lehet problémás. A törvénytervezet explicite fogalmaz, a termék kibocsátója nem védekezhet azzal, hogy egy esetleges rosszindulatú felhasználást nem látott előre. Véleményem szerint azonban abszurd helyzetet eredményezne az a követelmény, hogy a kibocsátó „előre kell lássa az előre láthatatlant". Olyan ez, mintha a kés gyártóján kérnénk számon, hogy egy késsel ölni is lehet – vagy még inkább (mert azt azért „ki lehetett volna találni") a fényképezőgép gyártóján számonkérni azt, hogy (és most szándékosan a gép egy meghökkentő használata következik) a finom masinával kalapálni is lehet.[5]

Egy vélemény szerint, ha az MI törvény a jelenlegi formájában kerül elfogadásra, megszűnik a „valódi" MI kutatás (pl. az egyetemeken), aminek a lényege, hogy félkész, ki tudja mire használható szoftvereket hoz létre és tesz közzé…

Szociális kreditrendszerek

Hogy az emberarcú MI mit jelent, azt egy másik projektből vett példával is jól illusztrálhatjuk. Elnyertünk erre egy 3 éves kutatási támogatást a VW Alapítványnál. A projekt (AI Fora) egyik tárgya a szociális kreditrendszerek vizsgálata – és megtiltása (Európában).

Szociális kreditrendszer alatt[6] átvitt értelemben minden olyan megoldást értenek, amely felfogásában és megvalósításában a kínai őspéldára hasonlít, és azzal rokon funkciókat valósít meg *(Mac Síthigh–Siems 2019)*. Az eredeti „Social Credit System" a Kínai Népköztársaság kormánya által kifejlesztett nemzeti hitelminősítő és feketelista. A program 2009-ben kezdte meg a regionális próbákat, majd 2014-ben nyolc hitelminősítő cég bevonásával országos kísérleti programot indított. Hivatalosan először az akkori kínai miniszterelnök, Wen Jiabao úr mutatta be 2011. október 20-án, az Államtanács egyik ülésén. A Nemzeti Fejlesztési és Reformbizottság (NDRC), a Kínai Népbank (PBOC) és a Legfelsőbb Népbíróság (SPC) által irányított rendszer célja a hitelminősítési funkció egységesítése, valamint a vállalkozások, kormányzati intézmények, magánszemélyek és nem kormányzati szervezetek pénzügyi és társadalmi értékelésének elvégzése.

A kreditrendszer szorosan kapcsolódik más kínai tömeges megfigyelési rendszerekhez, például a Skynethez, amely arcfelismerést, nagy adatok elemzését és különféle más MI alkalmazásokat foglal magában. (Nem igaz egyébként az a – véleményem szerint rasszista felhanggal – hangoztatott mondás, hogy „minden kínai egyforma", hiszen pl. működik rájuk az arcfelismerés.)

A rendszer támogatói mármost azt állítják, hogy az segít szabályozni a társadalmi viselkedést, javítja az állampolgárok „megbízhatóságát" (ami magában foglalja az adók és számlák időben történő befizetését), és elősegíti a hagyományos erkölcsi értékek elterjedését. A rendszer kritikusai szerint azonban az messze túllép a jogállamiságon és sérti az állampolgárok és a szervezetek törvényes jogait – különösen a jó hírnévhez, a magánélethez és a személyes méltósághoz való jogot, és hogy a rendszer a kormány átfogó felügyeletének és a Kínai Kommunista Párttal (KKP) szembeni ellenvélemények elnyomásának eszköze lehet. Különösen a különféle információforrások „egy kézben tartását" és integrációját kifogásolják. A szóban forgó aggályok mellett a szociális kreditrendszernek „nem tett jót" az sem, hogy a médiában nagy mennyiségű téves beszámoló és tévhit terjedt el a fordítási hibák, a szenzációhajhászás, az egymásnak ellentmondó információk és az átfogó elemzés hiánya miatt. A rossz példák közé tartozik az a széles körben elterjedt téves feltételezés, hogy a kínai polgárokat a rendszer által kiosztott numerikus pontszám alapján jutalmazzák és büntetik. Akárhogy is, azonban a szociális kreditrendszer alkalmas lehet arra, hogy a vonatjegyvásárlástól az egyszerű bolti fizetésig csak annak engedje az adott műveletet végrehajtani, aki „jó háttérrel" rendelkezik – például, akinek hitelkártyája van (és az adósságait rendben fizeti) –, vagy akár (ez is egy lehetőség) párttag. A rendszer „mindent" tud rólunk, a többi hamar megy. A rendszer legnagyobb problémája épp ezért nem is a jelenlegi használata, hanem az, hogy lehetőséget teremt a különféle visszaélésekre.

Ezt egy demonstrációs alkalmazással tudtuk bemutatni, amely egyszerűsített szociális kreditrendszert valósít meg. Rendelkezik egy „mindent" tartalmazó adatbázissal és egy szándékosan „gonoszra vett" tanuló rendszerrel.

A demonstráció során mesterségesen előállított személyes adatokat használtunk fel – hogy miért nem valódiakat, azt (különösen a GDPR hatálya alatt) talán már nem kell magyarázni. Fogtuk hát a valódi (természetesen anonimizált, azaz névtelen) adatokat, és véletlenszerűekkel helyettesítettük őket, úgy azonban, hogy az adatok eloszlása azonos maradjon. Így olyan adatokat kaptunk, amelyek nem voltak valódiak (mert egyetlen személy adataival sem egyeztek meg), de azok lehettek volna.[7] Nézzünk egy valódi, de azért mégis egyszerű példát! Mindenki valahány órát tölt naponta a munkahelyén – ha tört számokat is megengedünk, akkor egész szép eloszlásgörbét kapunk. Nos, a fentiek szerint mi úgy generál-

[5] „Legszebb álmaink is megvalósíthatók!" (Örkény I.) https://orkeny-egypercesek.blogspot.com/2012/01/legmereszebb-almaink-is-megvalosithatok.html.
[6] https://en.wikipedia.org/wiki/Social_Credit_System.
[7] Ehhez az ALLBUS (https://www.gesis.org/en/allbus/allbus-home) adatbázisát használtuk fel.

tunk egy-egy mesterséges adatot, hogy ennek a szóban forgó, az adott adatra (esetünkben a munkahelyen töltött időre) jellemző eloszlásgörbének megfeleljen. Ha e görbe maximuma mondjuk 7,3 óránál van, értéke pedig például 0,41 – ez azt jelenti, hogy a kérdőívet kitöltő személyek 41%-a 7,3 órát dolgozott – akkor ebből egy olyan mesterséges számot képeztünk, amely 41% eséllyel 7,3 lett (és hasonlóan a görbe többi pontjára, ezek együtt pedig természetesen 100%-ot adtak). Ilyen volt tehát a bemeneti információ, és ezután következett maga a kísérlet.

Ennek során a bemeneti információhalmazt (amelynek a munkahelyen töltött idő természetesen csupán egy eleme volt) egy tanuló rendszernek prezentáltuk. Tanítóval végzett tanulást végeztettünk a rendszerrel, és meghökkentő („gonosz") kérdésekre kerestük a választ: pl. hogy milyen eséllyel lesz az adott információhalmaz birtokosa a következő két évben elmebeteg (!) vagy követ el bűncselekményt (!). (Természetesen a tanító halmaz mindenféle esetet tartalmazott, azt is, hogy igen, meg azt is, hogy nem. A tanítás mindig az utólagos információra épült: a két év elteltével felvett adatokra. Akkor már tudni lehetett, ki lesz elmebeteg vagy bűnöző.)

Mi magunk döbbentünk meg a legjobban: közel 90% eséllyel meg lehetett jósolni a tesztelés során, hogy kiből lesz elmebeteg vagy bűnöző![8] (Nem mindenki lesz nagy eséllyel az, és sajnos nem tudtuk megállapítani, hogy ez min múlik – a tanuló rendszer egy híresen „átlátszatlan" NN volt… [ld. a következő pontot is]. Nem végeztünk olyan kísérletet, hogy mi történik, ha a – képzelt – személyekre vonatkozó adatbázis méretét, dimenzióját, azaz az eltárolt adatok számát módszeresen csökkentjük. Így feltehetően meg lehetett volna találni a „felelőst".)

Az egészből mindenesetre az látható, hogy egy „szép új világ" küszöbén állunk, amikor „az" MI többet tud rólunk, mint azt mi valaha is gondolnánk, és egy kellően rosszindulatú tanuló rendszer pedig mindenfélét megállapíthat, amiről nem is álmodtunk. (A határ a fantázia és a csillagos ég…)

Egyszóval van tárgya (és tétje) a Humane AI kutatásnak…

Megmagyarázható MI

A „megmagyarázható MI" (explainable AI vagy XAI, *Arrieta et al. 2020*) az utóbbi évek egyik slágertémája lett. Arra épül, hogy az MI alkalmazások többnyire „átlátszatlanok", homályosak: különösen igaz ez a tanuló rendszerekre és definíció szerint érvényes az idegháló- zat- (NN-) alapú tanuló rendszerekre (amelyek meg sem próbálnak érthetőek lenni). A másik jelenlegi „divat", a „deep learning", a mély tanulás általában amúgy is egy különösen átlátszatlan NN alkalmazását jelenti. Természetes, hogy nem tudni, mit miért csinál a rendszer. Nosza, változtassunk ezen! És máris megszületett a megmagyarázható MI. (Nem magyarázom, hogy miért „emberarcú MI" ez.)

A megmagyarázható MI (XAI) olyan MI, amelyben a megoldás eredményei az ember számára is érthetőek. Ez szemben áll a gépi tanulás „fekete doboz" fogalmával, ahol még a tervezők sem tudják megmagyarázni, hogy a mesterséges intelligencia miért is jutott egy adott döntésre. Az MI rendszerek felhasználói által használt privát mentális modellek alkalmazásával, finomításával és a téves elképzelések lebontásával az XAI azt ígéri, hogy segít a felhasználóknak hatékonyabb megértést nyújtani. Az XAI a magyarázathoz való társadalmi jog megvalósítása is lehet. Az XAI azonban akkor is lehet releváns, ha nincs törvényes jogi vagy szabályozási követelmény mögötte. Az XAI célja, hogy megmagyarázza, mit tettek, mit tesznek most, mit fognak tenni legközelebb, és feltárja a cselekvések alapjául szolgáló információkat. Ezek lehetővé teszik (i) a meglévő tudás megerősítését, (ii) a meglévő tudás megkérdőjelezését és (iii) az új feltételezések generálását.

Az XAI jellemzően vagy (i) kontrafaktuális – tényellentétes – elemzést nyújt vagy (ii) racionalizációt, észszerűnek mutatva be az MI meghozott döntését. Az (i)-re példa, ha a hitelkérelmet elutasították, és ahhoz, hogy osztályt váltsunk (azaz számunkra kedvezőbb döntés szülessen), az MI szerint több jövedelemre vagy több megtakarítási évre lenne (lett volna) szükség.

A terület a legnagyobb sikereit azonban mégis a (ii)-ben, nevezetesen éppen a tanuló rendszerek terén érte el: az ún. „tanítóval való tanulás" esetén valaki (ő a tanító) tudja a helyes megoldást, és visszacsatolást nyújt, visszajelzést adva a rendszernek, hogy mennyire volt „jó" az, amit csinált. (A többi a tanuló algoritmus – eljárás – dolga, amely a kurrens kimenetből és a visszacsatolásból kiszámítja a szükséges változtatás irányát és mértékét. Eleinte nagy lépésekkel halad – ha véletlenszerű, random helyzetből indulunk –, aztán a célhoz közeledve „tipegni" kezd.) Nos, az XAI ehhez annyit tesz hozzá, hogy a tanítás tárgya most már nemcsak az eredeti kimenetet létrehozó rendszer, hanem egy magyarázatot létrehozó is. Ha a magyarázat jó vagy rossz, a tanító fel fogja ismerni, és ennek megfelelően avatkozik be.

Az XAI nagy előrelépést ígér az MI-t (ha kell, ha nem) ténylegesen övező „titokzatosság" lebontásában.

Az Alfi projekt

Röviden ki szeretnék térni egy általunk fejlesztett alkalmazásra is *(Konok et al. 2021),* melynek első változata a beszédes nevű „Jó Mobil" OTKA projekt keretén belül készült, ma pedig az MI Nemzeti Laboratórium támogatását élvezi. Ez véleményem szerint az „emberarcú MI" iskolapéldája, és nagyobb potenciállal rendelkezik,

[8] A tesztelés során is rendelkezésre álltak a két évvel későbbi adatok. Főleg a bűnelkövetésnél lesz a dolog etikailag érdekes: ha mondjuk az jön ki, hogy nagy valószínűséggel bűnöző lesz valakiből, akkor le lehet tartóztatni, mielőtt még bármit elkövetett volna? (És mi van, ha mégsem követi el?) Szerencsére e nehéz kérdést nem nekem kell megválaszolni, de fel lehetett tenni.

mint az a kurrens ismertségéből következik (noha a hazai média többször is beszámolt róla[9]).

Az Alfi egy játékos alkalmazás gyerekeknek, ami segíti a digitális eszközök tudatosabb és egyben határok között tartott felhasználását, miközben különféle készségeket fejleszt. A kutatás egyik fő kérdése az volt, hogy valóban van-e különbség a digitális eszközöket sokat használó („kütyüzők") és az azokat egyáltalán nem használó gyerekek („nem kütyüzők") képességei közt, és hogy ez a különbség milyen területeket érint pontosan (pl. a kognitív, érzelmi, szociális fejlődést).

A kutatás másik fő kérdése, hogy hogyan lehet a digitális eszközöket készségfejlesztésre használni. A túlzásba vitt „kütyüzés" például nyilvánvalóan elvonja az időt a társas-kreatív tevékenységektől (ez eddig nem egy „nagy felfedezés", ez nem lehet másként), amitől ezeknek a készségeknek a fejlődése megrekedhet (ez viszont egy vizsgálandó és vizsgálható kérdés). A mobilalkalmazásunk segítségével igyekszünk ösztönözni a társas tevékenységeket, akár egy szülővel közösen olvasható interaktív mese segítségével, akár párosan játszható fejlesztő játékok által.

Műszakilag mindez pedig egy „kids launcher"-t (gyereknek szánt programindító alkalmazást) jelent, amit a kis felhasználó a „telefonnal" azonosít (a célközönség a 4–6 éves gyerekek!).

A rendszer egy „nagymama" funkcióit valósítja meg, „aki" érzelmeket mutat a gyermek felé: például dühös, szeret stb., ennek megfelelő hangokat ad ki (!), és valós szociális interakciókat facilitál (pl. elküldi a gyereket focizni!). Tartalmaz a létrejött megoldás számos, a használatot segítő és „vonzóvá tevő" eredeti elemet is, például interaktív mesét és sok egyebet. Van olyan funkció is benne, amit csak kettesben lehet megvalósítani a gyermek felügyelőjével, például édesanyjával, ezért reményeink szerint a gyermek felhasználó „nyaggatni" fogja őt, hogy érjen rá…

Zárszó és kitekintés

Az alábbiakban előbb összegzem a cikk főbb állításait, majd folytatom azokat.

Az írás a témában tartott előadáson alapul. Másik (német) munkahelyemen a „HumanE AI Net" nevű nagyszabású (12 millió euróval támogatott) uniós projekt menedzsere vagyok, amely 53 vezető uniós intézményt foglal magában, köztük nagy egyetemeket (UCL London, LMU München, Sorbonne, Sussex vagy ELTE), kutatóintézeti hálózatokat (Fraunhofer, Max Planck Gesellschaft, INRIA, CNR Olaszország), nagy nemzetközi vállalatokat (ING Bank, SAP, Philips, Airbus) stb. Az írásban a humán mesterséges intelligenciával kapcsolatos általános kérdéseket, a tervezett EU AI törvényt, a szociális hitelrendszereket, a megmagyarázható mesterséges intelligenciát (XAI), illetve az Alfi GK projektet tárgyalom.

Az Európai Bizottság 2021 áprilisában javaslatot tett a mesterséges intelligenciáról szóló rendeletre, az úgynevezett AI Act-re. A rendelet célja az emberarcú mesterséges intelligencia európai dimenzióban. Bár még csak tervezetről van szó, a tét igen nagy. A tervezett törvénynek pedig vannak hibái (ezt állítom), amelyeket ki kell javítani, mielőtt a szöveg törvényként átmegy.

Egy másik megvitatandó téma a szociális hitelrendszerek tanulmányozása – és tiltása (legalábbis Európában). Az eredeti „Social Credit System" a Kínai Népköztársaság kormánya által kifejlesztett nemzeti hitelminősítő és feketelista. A rendszer támogatói azt mondják, hogy az segíti szabályozni a társadalmi viselkedést, javítja a polgárok „megbízhatóságát" (ami magában foglalja az adók és számlák időben történő befizetését), és elősegíti a hagyományos erkölcsi értékek terjedését. A rendszer kritikusai azonban azzal érvelnek, hogy a rendszer messze túlmutat a jogállamiságon, és sérti az emberek törvényes jogait – különösen a jó hírnévhez, a magánélethez és a személyes méltósághoz való jogot –, valamint, hogy a rendszer a kormány kiterjedt megfigyelésének és az ellenvélemények elnyomásának eszköze lehet.

A „megmagyarázható mesterséges intelligencia" (XAI) az elmúlt években forró témává vált. Az AI-alkalmazások többnyire „átláthatatlanok": ez különösen igaz a tanuló rendszerekre és a neurális hálózatokra (NN). A jelenlegi divat, a „mélytanulás" általában amúgy is egy különösen átláthatatlan NN alkalmazását jelenti. Természetes, hogy nem tudjuk, mit és miért csinál a rendszer. Változtassunk tehát ezen! Ezzel a tétellel született meg az XAI. Áttekintek néhány megoldást a problémára.

Az írásban megemlítek egy alkalmazást is, az Alfi-t, amelynek első változata az OTKA „Jó mobil" projektben készült, és most az MI Nemzeti Laboratórium támogatja. Az Alfi egy tudományos alapú, játékos alkalmazás gyerekeknek, amely segít a digitális eszközök tudatosabb és korlátok között történő használatában, miközben számos készséget fejleszt. Egy „nagymama" funkcióit látja el, aki érzelmeket mutat a gyermek felé: lehet pl. dühös, szerető stb. Az alkalmazás a megfelelő hangokat (!) adja ki, és elősegíti a valódi szociális interakciókat (pl. elküldi a gyermeket focizni (!)).

Egy jó mondás („bon mot") szerint a ma számítástechnikája a tegnap MI kutatása.[10] És valóban: ha például a mobiltelefont vesszük, az tele van olyan szoftverekkel, amik ismert MI-problémákat „oldanak meg": persze néha a megoldás nem optimális a rendelkezésre álló kevés idő és a korlátos számítási erőforrás miatt. A példákat lehet sorolni az útvonaltervezéstől kezdve (amit valóban „nem lehet" egy telefonon rendesen megoldani: az utazó ügynök probléma közismert nehézségekhez vezet

[9] A projekt honlapja a https://www.alfageneracio.hu, ami a vonatkozó linkeket is listázza.

[10] https://en.wikipedia.org/wiki/History_of_artificial_intelligence

– egészen más azonban a helyzet, ha nem az optimális megoldást keressük, hanem *egy jó* megoldással is megelégszünk), a hangfelismerésen át a kézírás elektronikus szöveggé alakításáig. (Jegyezzük meg, hogy ezek többsége nem tanuló rendszerként lett megvalósítva, lásd az „MI törvény" kapcsán mondottakat.)

Ha a mondást igaznak tartjuk (mint ahogy…), akkor az is következik belőle, hogy a jelen MI kutatása a jövő számítástechnikája. Nem olyasféle speciális MI megoldások várhatók tehát, amelyek csak egy szűk kört érintenek (azokat, akik speciálisan az illető területen dolgoznak), hanem a mai MI kutatás eredményei a jövőben várhatóan széles körben elterjedt, alapértelmezett szoftverek lesznek, mint a jelenben a fent említett térkép vagy hangfelismerés. (Természetesen az viszont nem igaz, hogy *mindenből* lesz majd *valami:* van, amiből lesz, és van, amiből nem lesz. Egy régebbi példát említve: ki emlékszik már a tervezési algoritmusok [plannning algorithms] problémáira?[11])

Az MI-kutatás jelenleg változóban van. Régebben egyáltalán nem foglalkoztak azokkal a kérdésekkel, amelyek ma az előtérben állnak. Mindez egy ingát juttathat az eszünkbe, amely hol erre, hol meg arra leng ki. Amikor kileng, akkor „túlságosan". A lengésideje pedig években, évtizedekben mérhető. Mindenesetre az várható, hogy az inga (hacsak új lendületet nem kap valamitől) egyszer csak majd lecseng, és az inga „beáll középre". Ez a helyzet a GDPR-rel is; korábban egyáltalán nem törődtek a személyi adatokkal és azok védelmével sem. Most persze lehet, hogy „átestünk a ló túlsó oldalára": de a mondottak miatt ez nem biztos, hogy baj. Majd „rendbe jön" a dolog. Egészen hasonló a helyzet az „emberarcú MI"-vel is. Most egyszerre rájöttek (ha cinikusak vagyunk: ugyanazok, akik eddig az ellenkezőjét vallották), hogy az MI-t mi, emberek, csináljuk, és ez csak akkor éri meg, ha az nekünk hasznos, vagyis ha emberléptékű, emberarcú.

[11] Szinte hallom a felhördülést: hogyhogy ebből „nem lett semmi"? Gondolom, megjelent ugyanis egy összefoglaló publikáció, és ma is van, aki ezzel foglalkozik. Mindez, gondolom, így van. Az áttörés azonban, mértéktartóan fogalmazva, még várat magára…

Köszönetnyilvánítás

Az itt ismertetett kutatási eredmények az „MI Nemzeti Laboratórium" (RRF-2.3.1-21-2022-00004) részleges támogatásával születtek.

Irodalom

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. & Chatila, R. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, Vol. 58, pp. 82–115.

Cataleta, M. S. (2021) Humane Artificial Intelligence. The Fragility of Human Rights Facing AI. East West Center

Konok, V., Liszkai-Peres, K., Bunford, N., Ferdinandy, B., Jurányi, Z., Ujfalussy, D. J., Réti, Z., Pogány, Á., Kampis, G. & Miklósi, Á. (2021) Mobile use induces local attentional precedence and is associated with limited socio-cognitive skills in preschoolers. Computers in Human Behavior, Vol. 120, 106758.

Mac Síthigh, D., Siems, M. (2019) The Chinese social credit system: A model for other countries? The Modern Law Review, Vol. 82. No. 6. pp. 1034–1071.

Veale, M., Borgesius, F. Z. (2021) Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, Vol. 22. No. 4. pp. 97–112.