Az Európai Bizottság

MESTERSÉGES INTELLIGENCIÁVAL FOGLALKOZÓ MAGAS SZINTŰ SZAKÉRTŐI CSOPORTJA

A MEGBÍZHATÓ MESTERSÉGES INTELLIGENCIÁRA VONATKOZÓ ETIKAI IRÁNYMUTATÁS TERVEZETE

ÖSSZEFOGLALÓ

Munkadokumentum az érdekelt felekkel folytatott konzultációhoz

Brüsszel, 2018. december 18.

A MEGBÍZHATÓ MESTERSÉGES INTELLIGENCIÁRA VONATKOZÓ

ETIKAI IRÁNYMUTATÁS-TERVEZET

A mesterséges intelligenciával foglalkozó magas szintű szakértői csoport
A megbízható mesterséges intelligenciára vonatkozó etikai iránymutatás-tervezet

Európai Bizottság
Kommunikációs Főigazgatóság

Kapcsolattartó Nathalie Smuha – AI HLEG koordinátor
E-mail-cím CNECT-HLG-AI@ec.europa.eu

European Commission
B-1049 Bruxelles/Brussel

Dokumentum közzététele angolul: 2018. december 18.

Ezt a munkadokumentumot az AI HLEG (a mesterséges intelligenciával foglalkozó magas szintű szakértői csoport) – a csoport tagjainak egyes pontokkal kapcsolatos egyedi álláspontjának sérelme nélkül, valamint a dokumentum végleges változatának sérelme nélkül – fogalmazta meg. A dokumentum további munka tárgyát képezi, a végleges változat pedig 2019 márciusában, a mesterséges intelligenciával foglalkozó európai szövetség révén az érdekeltekkel folytatott konzultációt követően fog megjelenni.

A mesterséges intelligenciával foglalkozó magas szintű szakértői csoportról további információk érhetők el az interneten (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

ÖSSZEFOGLALÓ

Ez a munkadokumentum az Európai Bizottság mesterséges intelligenciával foglalkozó magas szintű szakértői csoportja (AI HLEG) által elkészített, a mesterséges intelligenciára vonatkozó etikai iránymutatás-tervezet, amelynek végleges változata 2019 márciusában fog megjelenni.

A mesterséges intelligencia (AI) korunk egyik legjelentősebb formáló ereje, amely várhatóan megváltoztatja társadalmunk szövetét. A mesterséges intelligencia nagyszerű lehetőséget biztosít a jólét és a növekedés fokozására, amelynek elérésére Európának is törekednie kell. Az elmúlt évtizedben az óriási mennyiségű digitális adat, a jelentős számítógépes architektúrák és az AI-technikák terén megvalósuló fejlesztések – mint például a gépi tanulás – rendelkezésre állásának köszönhetően jelentős előrelépésekre került sor. A jelentősebb AI-alapú fejlesztések az önvezető járművek, az egészségügy, a háztartási/szolgáltató robotok, az oktatás vagy a kiberbiztonság terén nap mint nap hozzájárulnak életminőségünk javulásához. A mesterséges intelligencia ráadásul a világ előtt álló számos jelentős kihívás megoldásában is kiemelt szerepet játszhat, mint például a globális egészség és jólét, az éghajlatváltozás, a megbízható jogi és demokratikus rendszerek, valamint egyéb, az ENSZ fenntartható fejlesztési céljai közé tartozó kihívások.

Azon túl, hogy a mesterséges intelligencia az egyének és a társadalom számára egyaránt hatalmas előnyöket jelenthet, bizonyos kockázatokat is hordoz magában, amelyeket megfelelően kell kezelni. Tekintettel arra, hogy egészében véve a mesterséges intelligencia által nyújtott előnyök meghaladják a vele járó kockázatokat, gondoskodnunk kell arról, hogy olyan úton haladjunk tovább, amely **maximalizálja a mesterséges intelligencia előnyeit, ugyanakkor minimálisra csökkenti a vele járó kockázatokat**. A helyes irány megtartása érdekében **a mesterséges intelligenciával kapcsolatos emberközpontú megközelítés szükséges**, amely annak észben tartására késztet, hogy a mesterséges intelligencia fejlesztésére és használatára nem szabad önmagában eszközként tekinteni, hanem olyan tényezőként, mint amelynek célja az emberi jólét növelése. **A megbízható mesterséges intelligencia lesz a mi vezércsillagunk**, hiszen az emberek csak akkor aknázhatják ki nyugodtan és teljes mértékben a mesterséges intelligencia előnyeit, ha megbízhatnak a technológiában.

A megbízható mesterséges intelligencia **két részből** áll: (1) tiszteletben kell tartania az alapvető jogokat, a vonatkozó szabályozást, valamint bizonyos alapelveket és értékeket, ami garantálja „**etikus célját**", továbbá (2) **technikai szempontból szilárdnak** és megbízhatónak kell lennie, mivel a technológia ismeretének hiánya még a jó szándék ellenére is okozhat kárt.

A jelen iránymutatás a **megbízható mesterséges intelligencia kereteit** határozza meg:

- Az **I. fejezet** a **mesterséges intelligencia etikus céljának biztosításával** foglalkozik, és megállapítja azokat az alapvető jogokat, elveket és értékeket, amelyeket a mesterséges intelligenciának be kell tartania.
- Ezekből az elvekből a **II. fejezet** levezeti a megbízható mesterséges intelligencia **megvalósításához nyújtott iránymutatást**, amely az etikus céllal és a technikai megbízhatósággal is foglalkozik. A fejezet felsorolja a megbízható mesterséges intelligencia követelményeit, és áttekintést nyújt azokról a műszaki és nem műszaki módszerekről, amelyek a megvalósítását szolgálják.
- A **III. fejezet** később ezeket a követelményeket a **gyakorlatba is átülteti** azáltal, hogy konkrét, de nem teljes körű értékelő listát állít fel a megbízható mesterséges intelligenciához. Ezt a listát a konkrét felhasználási területekhez kell igazítani.

Az etikus mesterséges intelligenciával foglalkozó egyéb dokumentumokkal ellentétben az iránymutatásnak nem célja, hogy a mesterséges intelligencia alapvető értékeinek és elveinek újabb listáját állítsa össze, hanem inkább az, hogy iránymutatást nyújtson ezen elvek és értékek konkrét

végrehajtásához és az AI-rendszerekben történő alkalmazásához. Az iránymutatás nyújtása három elvonatkoztatási rétegen keresztül történik: a legelvontabb az I. fejezetben olvasható (alapvető jogok, elvek és értékek), míg a legkonkrétabb a III. fejezetben (értékelő lista).

Az iránymutatás valamennyi, **a mesterséges intelligenciát fejlesztő, bevezető vagy használó érdekelt félnek** szól, beleértve a vállalatokat, szervezeteket, kutatókat, közszolgálatokat, intézményeket, magánszemélyeket vagy más érdekelt feleket. Az iránymutatás végleges változatában egy javasolt mechanizmus is szerepelni fog, amely lehetővé teszi az érdekelt felek számára az iránymutatás önkéntes elfogadását.

Fontos megjegyezni, hogy az iránymutatásnak nem célja, hogy a politikai döntéshozatal vagy a szabályozás bármely formájának a helyébe lépjen (ezzel az AI HLEG második, 2019 májusában esedékes munkaanyaga foglalkozik, melynek címe „A mesterséges intelligenciával kapcsolatos szakpolitikai és beruházási ajánlások"), vagy hogy gátolja azok megvalósítását. Az iránymutatást továbbá úgy kell tekinteni, mint egy olyan élő dokumentumot, amelyet rendszeres időközönként frissíteni kell, biztosítva ezáltal relevanciájának fenntartását, tekintettel a technológia és az azzal kapcsolatos ismereteink fejlődésére. Ezért e dokumentumnak kiindulópontként kell szolgálnia a „**Megbízható mesterséges intelligencia európai módra**" című vitához.

Miközben Európa csak akkor tudja a mesterséges intelligenciával kapcsolatos etikus hozzáállását közvetíteni, ha globális szinten is versenyképes, **a mesterséges intelligenciához való etikus hozzáállás kiemelten fontos a felelősségteljes versenyképesség elérése** szempontjából, hiszen erősíti a felhasználói bizalmat, és elősegíti a mesterséges intelligencia szélesebb körű használatát. Az iránymutatásnak nem az a célja, hogy gátolja Európában a mesterséges intelligenciával kapcsolatos innovációt, hanem sokkal inkább az, hogy az etikát ösztönzésként használja fel a mesterséges intelligencia sajátos európai védjegyének kialakításához, amely egyaránt szolgálja a magánszemélyek és a közjó javát és védelmét. Ez lehetővé teszi Európa számára, hogy az élvonalbeli, biztonságos és etikus mesterséges intelligencia piacvezetőjeként pozicionálja magát. Az európai polgárok csak a megbízhatóság garantálása mellett aknázhatják ki teljes mértékben a mesterséges intelligencia nyújtotta előnyöket.

Az iránymutatásnak emellett az is célja, hogy Európán túl is **elősegítse a közös gondolkodást és vitát** a mesterséges intelligencia **globális szintű**, etikus kereteiről.

IRÁNYMUTATÁS

Az iránymutatás minden egyes fejezete útmutatást nyújt a megbízható mesterséges intelligencia megvalósításához, és valamennyi, a mesterséges intelligenciát fejlesztő, bevezető vagy használó érdekelt félnek szól; lásd az alábbi összefoglalást:

I. fejezet: Kiemelt iránymutatás az etikus cél biztosításához:

- A mesterséges intelligencia **emberközpontúságának** biztosítása: a mesterséges intelligencia fejlesztése, bevezetése és alkalmazása „etikus céllal" kell történjen, amelynek alapjául az alapvető jogok, társadalmi értékek és a következő etikai alapelvek szolgálnak: *jótékonyság* (tégy jót!), *károkozás tilalma* (ne árts!), *az ember autonómiája*, *igazságosság* és *indokolhatóság*. Ez kiemelten fontos a **megbízható mesterséges intelligencia** eléréséhez.
- Az alapvető jogokra, etikai elvekre és értékekre kell támaszkodni, hogy fel tudjuk mérni a mesterséges intelligencia emberekre és a közjóra esetlegesen gyakorolt hatásait. **Különös figyelmet** kell fordítani a **veszélyeztetett csoportokat** érintő helyzetekre, mint például a gyermekek,

fogyatékossággal élők vagy kisebbségek, vagy **hatalmi illetve információs aszimmetriákkal** járó helyzetekre, például munkaadók és munkavállalók, vagy vállalkozások és fogyasztók között.
- Fel kell ismerni és tudomásul kell venni azt a tényt, hogy bár a mesterséges intelligencia komoly előnyöket nyújthat a magánszemélyeknek és a társadalomnak, kedvezőtlen hatással is járhat. Kiemelt figyelmet kell fordítani a kritikus alkalmazási területekre.

II. fejezet: Kiemelt iránymutatás a megbízható mesterséges intelligencia megvalósításához:

- A **megbízható mesterséges intelligencia követelményeit a tervezés legkorábbi fázisától** kezdődően be kell építeni: elszámoltathatóság, adatkormányzás, mindenki számára történő tervezés (Design for all), a mesterséges intelligencia autonómiájának irányítása (ember általi ellenőrzés), megkülönböztetésmentesség, az emberi autonómia tiszteletben tartása, a magánélet tiszteletben tartása, megbízhatóság, biztonság, átláthatóság.
- Figyelembe kell venni azokat a technikai és nem technikai módszereket, amelyek e követelményeknek az AI-rendszerben történő végrehajtását biztosítják. Továbbá e követelményeket akkor is szem előtt kell tartani, amikor felállítják a rendszeren dolgozó csapatot, magát a rendszert, a tesztkörnyezetet és a rendszer lehetséges alkalmazásait.
- Érthető és proaktív módon kell **tájékoztatni az érdekelt feleket** (ügyfeleket, munkavállalókat stb.) az AI-rendszer lehetőségeiről és korlátairól, ami lehetővé teszi számukra, hogy reális elvárásokat támasszanak a rendszerrel szemben. Az AI-rendszer **nyomonkövethetőségének** biztosítása kulcsfontosságú ebben a tekintetben.
- A megbízható mesterséges intelligenciát a **szervezeti kultúra részeként** kell kezelni, és tájékoztatni kell az érdekelt feleket arról, hogy a megbízható mesterséges intelligenciát hogyan valósítják meg az AI-rendszerek tervezése és használata során. A megbízható mesterséges intelligenciát a szervezet etikai kódexébe vagy magatartási kódexébe is be lehet illeszteni.
- Az AI-rendszer tervezése és fejlesztése során biztosítani kell az **érdekelt felek** részvételét és **bevonását**. A terméket fejlesztő, üzembe helyező és tesztelő csapatok összeállításánál pedig garantálni kell a **sokszínűséget**.
- Törekedni kell az AI-rendszerek **ellenőrizhetőségének elősegítésére**, különösen kritikus kontextusokban és helyzetekben. A rendszereket – amennyire csak lehetséges – úgy kell kialakítani, hogy lehetővé tegyék az egyes döntések különböző inputokig, adatokig, előre betanított modellekig stb. történő visszakövetését. Ezenfelül meg kell határozni az AI-rendszer **magyarázó módszereit**.
- Külön eljárást kell kidolgozni az **elszámoltathatóság irányításához**.
- Gondoskodni kell a **képzésről és oktatásról**, és biztosítani kell, hogy a vezetők, fejlesztők, felhasználók és munkaadók ismerjék a megbízható mesterséges intelligenciát, és képezzék magukat e területen.
- Ne feledjük, hogy jelentős feszültség alakulhat ki a különféle célkitűzések között (az átláthatóság megnyithatja az utat a visszaélések előtt; az elfogultság azonosítása és kiküszöbölése ellentétben állhat az adatvédelmi szempontokkal). Ezeket a kompromisszumokat nyilvánosságra kell hozni és dokumentálni kell.
- Elő kell segíteni a kutatást és innovációt, amelyek révén könnyebben megvalósíthatók a megbízható mesterséges intelligencia követelményei.

III. fejezet: Kiemelt iránymutatás a megbízható mesterséges intelligencia értékeléséhez

- A mesterséges intelligencia fejlesztése, bevezetése vagy használata esetén el kell fogadni a megbízható mesterséges intelligencia **értékelő listáját**, és azt ahhoz a konkrét felhasználási területhez kell igazítani, amelyen az adott rendszert használják.
- Ne feledjük, hogy az értékelő lista **soha sem lehet teljes körű**, és hogy a megbízható mesterséges

intelligencia biztosítása nem a szövegdobozok kipipálásáról szól, hanem sokkal inkább a követelmények azonosításának, a megoldások értékelésének és a jobb eredmények biztosításának állandó folyamatáról, amely az AI-rendszer teljes életciklusára kiterjed.

Ez az útmutatás egy, a mesterséges intelligencia emberközpontú megközelítését magáénak valló elképzelés részét képezi, amely lehetővé teszi Európa számára, hogy az etikus, biztonságos és élvonalbeli mesterséges intelligencia terén a világ vezető innovátora legyen. Célja, hogy elősegítse és lehetővé tegye az „**Európában készült megbízható mesterséges intelligenciát**”, amely fokozza az európai polgárok jólétét.