Ernst & Young Kft.
Ernst & Young Ltd.
H-1132 Budapest Váci út 20.
1399 Budapest 62. Pf.632, Hungary

Tel: +36 1 451 8100
Fax: +36 1 451 8199
www.ey.com/hu
Cg. 01-09-267553

Schrettner Attila

AI Ethics - 4. rész

Hogyan kontrollálható a mesterséges intelligencia (AI)?

** AI specifikus kontrolltevékenységek**

Ez a cikk az AI Ethics téma feldolgozásának lezáró cikke, a sorozat előző részei elérhetők a következő linkeken: Bevezetés, Data governance, Kockázatkezelés. Az egyes témák a kapcsolódó EU direktíva három nagyobb elvárás csoportja köré szerveződnek: data governance, kockázatkezelés és AI-specifikus kontrolltevékenységek (jelenlegi cikk)

I. A mesterséges intelligencia kockázatainak kezelése

A kontrollok megvalósításánál a kockázatalapú megközelítést érdemes követni az EU rendelet tervezet[1] szellemiségével is összhangban. Azaz az adott szervezet által használt AI rendszereket az előző cikkben tárgyalt kockázatok felmerülése mentén kockázati kategóriákba érdemes sorolni, és azokhoz rendelni a kontrolltevékenységeket. A három, az AI rendszerekre specifikus területet érdemes megismerni a megfelelő kontrollhoz:

- Emberi felügyelet („*human oversight*") – Az AI rendszer működésébe milyen módon történik emberi beavatkozás
- „Leiratkozás" („*opt-out*") biztosítása – A mesterséges intelligencia használata/hatálya alól milyen szempontok mentén kapnak felmentést/felmentési lehetőséget az adott szervezet partnerei, ügyfelei stb.
- Magyarázhatóság („*explainability*") garantálása – Az AI által generált eredmények hogyan magyarázhatók el/mutathatók be egyszerűen és személetesen az alkalmazó szervezet döntéshozói, valamint a rendszer felhasználói vagy más érintettek számára.

Ezek persze nem az alkalmazható/alkalmazandó kontrolltevékenységek teljes körét jelentik (lehetnek még pl.: IT biztonság, adatvédelem, AI ethics board, AI desing standardok[2]). Azonban ezek azok a területek, amelyek az AI sajátosságaira reflektálnak és más rendszerek esetén nem feltétlenül, vagy nem ilyen formában jelennek meg. Fontos még az oktatás és képzés szerepe is, hogy a mesterséges intelligenciával dolgozó munkavállók tisztában legyenek a mögöttes logikával és az általuk alkalmazott matematikai alapokkal, mint például a lineáris algebra[3].

[1] Az Európai Parlament és a Tanács Rendelete A Mesterséges Intelligenciára Vonatkozó Harmonizált Szabályok (A Mesterséges Intelligenciáról Szóló Jogszabály) Megállapításáról és Egyes Uniós Jogalkotási Aktusok Módosításáról
[2] EY (2018): How do you teach AI the value of trust?
[3] Saniya Parveez, Roberto Iriondo (2021): Basic Linear Algebra for Deep Learning and Machine Learning Python Tutorial, An introductory tutorial to linear algebra for machine learning (ML) and deep learning with sample code implementations in Python, Towards AI, Link: https://pub.towardsai.net/basic-linear-algebra-for-deep-learning-and-machine-learning-ml-python-tutorial-444e23db3e9e

II. Emberi felügyelet a mesterséges intelligencia felett

A kockázatalapú megközelítés szerint érdemes meghatározni az emberi felügyelet mértékét is. Alapvető, hogy minél magasabb a kockázat annál jelentősebb humán kontrollt javasolt alkalmazni. Ezen felül persze további fontos szempontokat is érdemes mérlegelni.

Az emberi felügyeletet három fő kategóriába sorolhatjuk[4], fentről lefelé erősségi sorrendben:

- *Human-in-the loop*: Ez a megközelítés, amikor az emberi részvétel a döntéshozatalban folyamatosan jelen van, Az AI rendszer inkább csak inputokat ad a végső döntésekhez. Ezt olyan esetekben célszerű alkalmazni, amikor hatékonyan beépíthető a folyamatba vagy az adott döntés meghozatalához elengedhetetlen humán szempontok mérlegelése is. Ilyenek például a mesterséges intelligencia hadászati alkalmazásai.

- *Human-over-the-loop*: Ebben a megközelítésben az emberi döntéshozó valamilyen logika mentén kiválasztott AI rendszer által hozott döntéseket vizsgál meg jóváhagyás előtt. Ez a kiválasztási logika kötődhet például statisztikai bizonyossághoz, valamilyen értékhatárhoz vagy bizonyos döntésekhez is.

- *Human-out-the-loop*: Ebben a megközelítésben az emberi részvétel a döntéshozatalban csak esetleges, vagy csak utólag kerül visszaellenőrzésre az AI rendszer működése. Rendszerint valamilyen gyakorlati tényező, például a mesterséges intelligencia által hozott döntések számossága miatt van. Ilyen lehet egy tartalomajánló algoritmus. Ebben az esetben kap igazán nagy jelentőséget az előző cikkekben tárgyalt *data governance* és kockázatkezelési rendszer.
-

Ezen felül léteznek más megközelítések is (*human-on-the-loop*, *human-in-command*, stb.) de ezek tárgyalása meghaladja a cikk kereteit[5].

III. „Leiratkozás" biztosítása az AI rendszerből

A leiratkozás („opt-out") lehetősége azt jelenti, hogy az AI rendszer felhasználóinak van rá lehetősége, hogy saját magukat bizonyos esetekben kivonják a mesterséges intelligenciahasználata/hatálya alól és lehetőséget kapjanak, hogy ügyüket más elbírálási módszer szerint ítéljék meg[6].

Ez természetesen nem jelenti azt, hogy egy szervezetnek minden partnerének, ügyfelének stb. és minden AI rendszer esetén kötelező lenne ezt a lehetőséget fenntartani. A lényeg, hogy magas kockázatú és monopolisztikus vagy kevés alternatívával rendelkező esetekben meglegyen a választás lehetősége. Például egy zenei streaming szolgáltató tartalomajánló algoritmusa esetén nem biztos, hogy indokolt ezt alkalmazni, de egy banki hitelelbírálás esetén már megfontolható.

Amennyiben a "leiratkozási" lehetőséget a kockázatértékelés alapján szükségesnek ítéli meg az adott szervezet, úgy azt a partnerek számára is könnyen hozzáférhető és közérthető módon kell kialakítani azt.

[4] World Economic Forum (2020): Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations
[5] EU High-Level Expert Group On Artificial Intelligence (2019): Ethics Guidelines For Trustworthy AI
[6] AI Ethics Impact Group (2020): From Principles to Practice: An interdisciplinary framework to operationalise AI ethics

Building a better working world

IV. Az AI magyarázhatóságának garantálása

A mesterséges intelligencia sokszor nem csak a felhasználók számára, hanem a velük dolgozó munkavállalók számára is „fekete dobozként" viselkedhet[7], azaz az AI rendszerbe betáplált adatok és az általa hozott döntések között nehéz mélyebb elemzés nélkül közvetlen kapcsolatot kimutatni. Külön kihívás lehet ez, ha az algoritmus több döntési réteget is tartalmaz (pl.: *neural network*).

Az AI döntéseink magyarázhatóságát három dimenzióra lehet felbontani[8]. Ezek fentről lefelé, a technikaitól az általános felé haladva:

- Mérnöki magyarázhatóság („*engineers' interpretability*"): A mögöttes matematikai modell érthetősége
- Alapvető magyarázhatóság („*casual interpretability*"): Az AI rendszer inputjai és outputjai közti összefüggés/kapcsolat érthetősége
- Bizalmi magyarázhatóság („*trust-inducing interpretability*"): Egyéb addicionális információk rendelkezésre bocsájtása szemléletes formában, amely segít a „laikus" felhasználóknak is a működés megértésében.

Nincs szükség feltétlenül mindhárom szintre, de a rendszert alkalmazó szervezetnek érdemes meghatároznia, hogy melyik érintett számára, melyik magyarázhatósági dimenzió(k) mentén kommunikál.

V. EU direktíva tervezet az AI specifikus kontrollokról

Az EU rendelet is hasonló elvárásokat fogalmaz meg a magas kockázatú AI rendszerek esetén[9]. Megjelenik az emberi felügyelet és a magyarázhatóság. A „leiratkozás"-t is értékeli a direktíva. Amennyiben ugyanis nincs „opt-out" lehetőség, az AI rendszer potenciális kockázati besorolását magasabb irányba toló tényezőként kell értelmezni.

Az emberi felügyelet[10] kapcsán alapelvárás, hogy az AI rendelkezzen ember-gép interfésszel, azaz a felügyelő személy és a mesterséges intelligencia között ki legyen alakítva a hatékony felügyeletet lehetővé tévő kezelő felület. Ezen felül biztosítani kell a humán kontrollt ellátó személyek számára, hogy a körülményeknek megfelelően:

- Értik a nagy kockázatú rendszer kapacitásait és korlátait;
- Képesek megfelelően nyomon követni a rendszer működését annak érdekében, hogy a rendellenességek, zavarok és váratlan teljesítmény jeleit a lehető leghamarabb fel tudják tárni és kezelni;
- Tudatában vannak, ha automatikusan vagy túlzott mértékben támaszkodnak a rendszer által előállított outputra („automatizálási torzítás");

[7] Ioannis Kakogeorgioua, Konstantinos Karantzalos (2021): Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing, International Journal of Applied Earth Observation and Geoinformation
[8] Katharine Miller (2021): Should AI Models Be Explainable? That depends., Stanford University, Stanford Institute for Human-Centered Artificial Intelligence.
[9] Az Európai Parlament és a Tanács Rendelete A Mesterséges Intelligenciára Vonatkozó Harmonizált Szabályok (A Mesterséges Intelligenciáról Szóló Jogszabály) Megállapításáról és Egyes Uniós Jogalkotási Aktusok Módosításáról
[10] Az Európai Parlament és a Tanács Rendelete A Mesterséges Intelligenciára Vonatkozó Harmonizált Szabályok (A Mesterséges Intelligenciáról Szóló Jogszabály) Megállapításáról és Egyes Uniós Jogalkotási Aktusok Módosításáról 2. fejezet, 14. cikk

- Képesek legyenek a rendszer outputjainak helyes értelmezésére;
- Képesek legyenek arra, hogy úgy dönthessenek, hogy nem használják a rendszert vagy más módon figyelmen kívül hagyják, felülírják vagy visszafordítsák annak outputjait;
- képesek legyenek beavatkozni a rendszer működésébe, vagy megszakítani a rendszert valamilyen eljárás segítségével.

A magyarázhatóság területén elvárás a megfelelő műszaki dokumentáció, nyilvántartás[11] (ezek inkább a mérnöki magyarázhatósághoz sorolhatók) a felhasználók tájékoztatása[12] (ez inkább a bizalmi magyarázhatósághoz sorolható) és az átláthatóság. Utóbbi tekintetében a következő információkat kell legalább bemutatni:

- A meghatalmazott képviselő kiléte és elérhetőségei;
- Az AI rendszer jellemzői, képességei és teljesítményének korlátai, beleértve a következőket:
    - Rendeltetés;
    - Pontosság, stabilitás és kiberbiztonság várható szintje;
    - A rendeltetésszerű használattal vagy az észszerűen előrelátható rendellenes használattal összefüggő kockázatok;
    - Az AI rendszer teljesítménye azon személyek vagy személyek csoportjai tekintetében, akikre vagy amelyekre használni kívánják;
    - A bemeneti adatokra vonatkozó előírások;
- Teljesítményt érintő változások (ha vannak ilyenek);
- Az emberi felügyeleti intézkedések;
- Várható életciklus, valamint a megfelelő működés biztosításához szükséges karbantartási intézkedések, beleértve a szoftverfrissítéseket is.

Ezen felül a rendelet megfogalmaz még pontossági, stabilitási és kiberbiztonsági elvárásokat is, azonban ezek nem vagy nem feltétlenül az AI rendszerekere specifikusak, így ezeket nem tárgyaljuk jelen cikkben.

VI. Emberi felügyelet és kapcsolódó belső ellenőrzési vizsgálati pontok

A fenti kontrolltevékenységeket mind vizsgálhatja a belső ellenőrzés. Most példaként az emberi felügyeletre fókuszálunk. Az alábbi területekre érdemes hangsúlyt fektetni[13]:

- Megfelelő a kockázati kategória besorolása?

[11] Az Európai Parlament és a Tanács Rendelete A Mesterséges Intelligenciára Vonatkozó Harmonizált Szabályok (A Mesterséges Intelligenciáról Szóló Jogszabály) Megállapításáról és Egyes Uniós Jogalkotási Aktusok Módosításáról 2. fejezet, 11, 12. cikk
[12] Az Európai Parlament és a Tanács Rendelete A Mesterséges Intelligenciára Vonatkozó Harmonizált Szabályok (A Mesterséges Intelligenciáról Szóló Jogszabály) Megállapításáról és Egyes Uniós Jogalkotási Aktusok Módosításáról 2. fejezet, 13. cikk
[13] Information Commissioner's Office (2020): Guidance on the AI auditing framework Draft guidance for consultation

- Helyesen határozták meg az emberi felügyelet módját?
- Milyen az ember-gép interfész működése?
- Szabályszerűen vezetik a műszaki dokumentációt és az egyéb kapcsolódó nyilvántartásokat?
- Megfelelő az AI által hozott döntések felülvizsgálatának átfutási ideje és minősége?

VII. Összefoglalás

A mesterséges intelligencia alkalmazása számos kockázatot hordoz ezért ezen a területen kiemelkedően fontos a megfelelő, jól felépített, az üzletet támogató, nem túl szabályozó és valós veszélyeket kezelő kontrollok alkalmazása. A cikk azokat a lehetőségeket mutatja be az EU vonatkozó irányelvével összhangban, amelyekkel megfelelően kezelhetőek az AI alkalmazásának kockázatai.