HADMÉRNÖK
VÉDELEMINFORMATIKA

19. évfolyam (2024) 3. szám 119–135.
DOI: 10.32567/hm.2024.3.8

Bederna Zsolt[1]

A mesterségesintelligencia-rendszerek megfelelősége

Compliance of Artificial Intelligence Systems

Absztrakt

A különféle mesterségesintelligencia- (MI) alapú megoldások terjedése következtében elengedhetetlenné vált az MI-rendszerek által jelentett kockázatok megértése és menedzselése. A tanulmány szisztematikusan vizsgálja az MI által képviselt kockázati profilokat, hangsúlyozva a személyre szabott kockázatkezelési stratégiák és etikai megközelítések jelentőségét.
Az elemzés feltárja a magas kockázatú MI-rendszerek kihívásainak jellegét, amelyek szigorú szabályozási megfelelést tesznek szükségessé a lehetséges káros következmények megakadályozása, illetve mérséklése érdekében, míg a jogalkotó ösztönzi a vonatkozó kötelezettségek teljesítését az alacsonyabb kockázatú MI-rendszerek esetén az átláthatóság és az elszámoltathatóság vonatkozásában.
Jelen tanulmány részletes áttekintést nyújt az MI kockázati tényezőiről, amelyek befolyásolják az adatok integritását, a modellek pontosságát, a folyamatok és eredmények megbízhatóságát, valamint a felhasználói interakciót.

Kulcsszavak: mesterséges intelligencia, Európai Unió, kockázatok, irányítási rendszerek, kiberbiztonság

Abstract

Due to the proliferation of various artificial intelligence (AI) solutions, understanding and managing the risks posed by AI systems has become essential. This study systematically

[1] Nemzeti Közszolgálati Egyetem Államtudományi és Nemzetközi Tanulmányok Kar Kiberbiztonsági Tanszék, e-mail: bederna.zsolt@bederna.hu

examines the risk profiles represented by AI, emphasizing the importance of personalized risk management strategies and ethical approaches.
The analysis reveals the nature of challenges posed by high-risk AI systems, which require strict regulatory compliance to prevent or mitigate potential harmful consequences, while lawmakers encourage the fulfilment of relevant obligations regarding transparency and accountability for lower-risk AI systems. The AI legislation mandates that providers of high-risk AI systems establish and operate an AI governance system that integrates quality management, information security, and data protection.
The paper provides a detailed overview of AI risk factors that affect data integrity, model accuracy, the reliability of processes and outcomes, as well as user interaction.

Keywords: artificial intelligence, European Union, risks, management systems, cybersecurity

Bevezetés

A mesterséges intelligencia (MI) múltját jellemző fellendülések és hullámvölgyek változó természete ellenére a jelenlegi trendek azt mutatják, hogy az MI-t az élet egyre több területén alkalmazzák, mint például az ipar, a gyártás, az egészségügy, az adatelemzés. Ennek oka az MI képességbeli és teljesítményét érintő fejlődése, amelynek alapját a mögöttes elmélet és algoritmus, valamint az információs és kommunikációs technológia (IKT) számítási kapacitásának fejlődése képezi.[2]

Az Európai Unió a mesterséges intelligenciát a digitális stratégiája részeként szabályozza azzal a céllal, hogy jobb feltételeket biztosítson a technológia fejlesztéséhez és használatához, illetve csökkentse az MI-rendszerek által jelentett kockázatokat. Az Európai Bizottság 2022-ben terjesztette elő az Európa digitális évtizede szakpolitikai programját,[3] amely konkrét célokat és célkitűzéseket tartalmaz 2030-ra az EU egyik prioritásának számító digitális átalakulással kapcsolatos területeken. A program magában foglalja (1) a digitális készségekbe való befektetést Európa új digitális technológiákkal kapcsolatos kapacitásainak megerősítése érdekében, (2) az emberek kiberfenyegetésekkel szembeni védelmét, a kiberbiztonsági szint és a kapcsolódó képességek javítását, (3) az ultragyors széles sávú internet elterjedésének felgyorsítását, (4) Európa szuperszámítógép-kapacitásának bővítését az orvostudomány, a közlekedés és a környezetvédelem terén történő innovatív megoldások kidolgozása érdekében és (5) annak biztosítását, hogy az MI fejlesztése az emberek jogainak tiszteletben tartásával történjen.

2021 áprilisában az Európai Bizottság javaslatot tett az MI szabályozási keretére.[4] A Mesterséges intelligenciáról szóló jogszabály[5] javaslatát várhatóan 2024 harmadik negyedévében fogadják el. Uniós szinten létrejött a tagállamok és a Bizottság képviselőiből álló Mesterséges Intelligenciával Foglalkozó Európai Hivatal, amely összegyűjti és megosztja a legjobb gyakorlatokat a tagállamok között. Nemzeti

[2] Jiang et al. 2022.
[3] *Az Európai Parlament és a Tanács (EU) 2022/2481 határozata.*
[4] Council of the European Union 2024.
[5] Európai Parlament 2024.

szinten a tagállamoknak ki kell jelölniük egy vagy több nemzeti illetékes hatóságot, amelyek közül a nemzeti felügyeleti hatóság felügyeli a jogszabály alkalmazását és a végrehajtását.

Jelen tanulmánnyal a szerző az MI-rendszerek ellenében az Európai Unió által megfogalmazott jogszabályi kötelezettségek, valamint az MI-rendszerek által jelentett és a működésüket jellemző kockázatok áttekintését és elemzését tűzi ki célul.

E célok elérése érdekében a szerző előbb a módszertant ismerteti, majd áttekinti a mesterséges intelligencia kockázati profiljait, a kötelezettségeket és a vonatkozó irányítási rendszereket. Ezután az MI-rendszerekkel összefüggésbe hozható információbiztonsági, adatvédelmi, illetve minőségi elvárásokat, a kapcsolódó irányítási rendszerek összefüggéseit elemzi. Továbbá az MI-rendszerek által jelentett, valamint az MI-rendszerekre hatással lévő kockázatokat vizsgálja. A cikk összegzéssel és konklúzióval zárul.

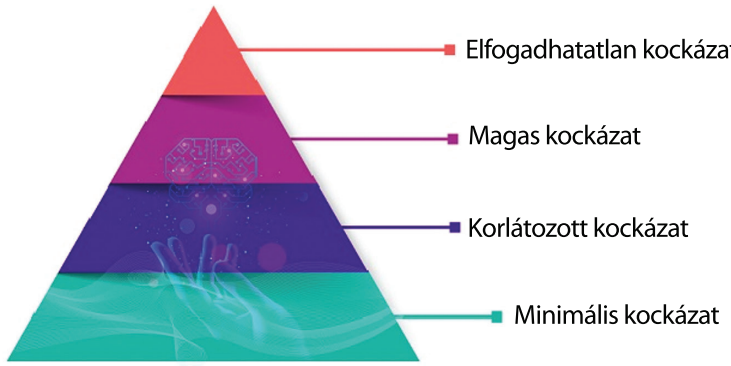
Módszertan

Elsődlegesen az MI-vonatkozású európai uniós jogszabályok, nemzetközi szabványok, illetve a hazai és nemzetközi releváns szakirodalom feldolgozásával azonosítjuk és rendszerezzük a jogszabályi kötelezettségeket. A jogalkotó által a Mesterséges intelligenciáról szóló jogszabályban megfogalmazott elvárt kötelezettségek és az ISO/IEC 42001 szabvány tekintetében összehasonlító elemzést végzünk annak érdekében, hogy azonosítani tudjuk más szakterületekkel (úgymint adatvédelem és kiberbiztonság) a kapcsolódási pontokat, valamint az elvárt kockázatkezelési megközelítést. Végezetül az MI-rendszerek által jelentett és a működésüket jellemző kockázatok átfogó elemzését, kategorizálását valósítjuk meg.

Jogszabályi áttekintés

Mesterséges intelligencia kockázati profilok

A Mesterséges intelligenciáról szóló jogszabály kimondja, hogy a különböző alkalmazásokban használható MI-rendszereket négy kockázati osztályba kell besorolni aszerint, hogy milyen kockázatot jelentenek a felhasználók számára. A jogalkotó négy kockázati szintet határozott meg: (1) minimális kockázat, (2) korlátozott kockázat, (3) magas kockázat és (4) elfogadhatatlan kockázat (1. ábra).



1. ábra: Mesterséges intelligencia kockázati profilok
Forrás: a szerző szerkesztése European Commission 2024 alapján

Minimális és alacsony kockázati profil

A jogszabály megfogalmazása szerint minimális kockázatot azok az MI-rendszerek jelentenek, amelyek nem tartoznak magasabb biztonsági osztályba. E rendszerek alkalmazására nincs korlátozás vagy kötelező tevékenység, mindenesetre javasolt az általános elvek követése, mint az emberi felügyelet, az egyenlő bánásmód és a méltányosság.

Alacsony kockázat esetén felmerül az MI-rendszerek által kivitelezhető manipuláció vagy megtévesztés lehetősége. Az általános célú MI (*general purpose artificial intelligence,* GPAI) nagy mennyiségű adatot felhasználva képes önálló döntések megalkotására, különböző feladatok kompetens végrehajtására. Egy GPAI-t önállóan vagy más rendszerbe vagy alkalmazásba integrált módon is kiadhatnak.

A GPAI-rendszerek kapcsán biztosítani kell a transzparenciát, azaz egyértelművé kell tenni a felhasználók számára, hogy MI-rendszerrel vagy MI-rendszer által létrehozott tartalommal dolgoznak (hacsak ez nem nyilvánvaló). Ennek megfelelően az MI segítségével generált vagy módosított tartalmat, például képeket, hang- vagy videófájlokat egyértelműen jelölni kell. Ezeknek a rendszereknek meg kell felelniük a digitális egységes piacon a szerzői és szomszédos jogokról szóló irányelvnek,[6] valamint meg kell akadályozni az illegális tartalom létrehozását.

Ezenkívül a GPAI-modellek valamennyi szolgáltatójának műszaki dokumentációt kell biztosítania, beleértve a képzési és tesztelési folyamatokat, valamint az értékelési eredményeket, továbbá azokat az információkat, amelyek a más szolgáltatók számára nyújtanak információt a GPAI-modell saját MI-rendszerbe történő integrálásához. Az ingyenes és nyílt licencű GPAI-rendszerek abban az esetben tartoznak a jogszabály hatálya alá, amennyiben nagy kockázatú MI-rendszerek részét képezik, illetve értelemszerűen ugyanúgy vonatkoznak rájuk a tiltott gyakorlatokkal kapcsolatos előírások és az átláthatósági kötelezettségi elvárások. E kategóriába a GPAI akkor tartozik, amennyiben a betanításhoz felhasznált számítások kumulatív mennyisége meghaladja

[6] Európai Parlament és a Tanács (EU) 2019/790 irányelve.

a 10^{25} lebegőpontos műveletet. A rendszerkockázattal járó GPAI-modellek szolgáltatóinak a korábbi kötelezettségeken túl a modellértékeléseket is el kell végezniük, továbbá megfelelő szintű kiberbiztonsági védelmet kell kialakítaniuk és fenntartaniuk, az incidensekről és az esetleges korrekciós intézkedésekről indokolatlan késedelem nélkül tájékoztatni kell az illetékes nemzeti hatóságokat.

Magas kockázati profil

A magas kockázatú MI-rendszerek negatívan befolyásolhatják egy természetes személy vagy egy csoport alapvető jogait, egészségét és biztonságát vagy a környezetet. A jogszabály III. melléklete szerint magas kockázatú MI-rendszernek minősülnek a) a biometrikus és biometria-alapú rendszerek; b) a kritikus infrastruktúra kezelésében és működtetésében közreműködő rendszerek; c) oktatást és szakképzést megvalósító rendszerek; d) foglalkoztatás és munkavállalók kezelését biztosító rendszerek; e) alapvető magán- és közszolgáltatásokhoz és juttatásokhoz való hozzáférést nyújtó rendszerek; f) bűnüldözést támogató rendszerek; g) migráció, menedékjog és határigazgatás kezelésében részt vevő rendszerek, valamint h) igazságszolgáltatást és demokratikus folyamatokat támogató rendszerek. Azonban kivételnek számítanak azok az MI-rendszerek, amelyek szűk feladatot látnak el, céljuk javítani egy korábban elvégzett emberi tevékenység eredményét, illetve döntéshozatali mintákat vagy eltéréseket észlelnek korábbi döntésekhez képest. További feltételként jelenik meg, hogy az MI-rendszer nem helyettesíti vagy befolyásolja a korábban elvégzett emberi értékelést megfelelő emberi felülvizsgálat nélkül, vagy mindössze előkészítő feladatot lát el.

Ahhoz, hogy az unió egységes piacán forgalomba vagy üzembe kerüljön egy magas kockázati besorolással bíró MI-rendszer, a szolgáltatónak meg kell felelnie a 8–25. cikkekben meghatározott követelményeknek. Ezek a kötelezettségek megkövetelik a mindenkori MI-szolgáltatótól, hogy hozzon létre kockázatkezelési és minőségirányítási rendszert, valamint megfelelő adatvédelmi megoldást, biztosítva, hogy a képzési, validációs és tesztelési adathalmazok relevánsak, megfelelően reprezentatívak és lehetőség szerint hibamentesek. Ezzel összefüggésben a kötelezően készítendő műszaki dokumentáció képes bizonyítani a rendszer megfelelőségét. Az ilyen rendszert úgy kell kialakítani, hogy a felhasználók emberi felügyeletet tudjanak megvalósítani, és elérjék a megfelelő mértékű pontosságot, robusztusságot és kiberbiztonságot.

A jogszabály további szereplőkre is kötelezettségeket ró. A magas kockázatú MI-rendszerek felhasználóinak megfelelő technikai és szervezeti intézkedéseket kell tenniük annak érdekében, hogy az ilyen rendszereket a rendelkezésre álló használati utasításoknak megfelelően használják (26. cikk). Az importőröknek ellenőrizniük kell, hogy (1) a szolgáltató elvégezte az adekvát megfelelőségértékelési eljárást, (2) a műszaki dokumentáció elkészült, (3) a rendszer viseli a szükséges CE-jelölést, és mellékelve van az EU megfelelőségi nyilatkozata, valamint (4) a szolgáltató kinevezett egy meghatalmazott képviselőt (23. cikk). Még a forgalmazók (24. cikk) és az értéklánc más szereplői (25. cikk) is felelősek azért, hogy ellenőrizzék és biztosítsák, hogy a szolgáltatók elvégezték a jogi kötelezettségeiket.

Nem elfogadható kockázati profil

A nem elfogadható kockázat a legmagasabb kockázati szint, amelyet a II. fejezet 5. cikke határoz meg. Értelemszerűen az e kategóriába tartozó tevékenységek tiltottak az EU értékeivel és alapvető jogaival való összeegyeztethetetlenség miatt. Ezek az alkalmazások a következőkhöz kapcsolódnak:

- Szubliminális technikák, amelyek túlmutatnak egy személy tudatosságán vagy célzottan manipulatívak vagy megtévesztők, és jelentősen rontják egy személy vagy csoport döntéshozatali képességét, aminek eredményeként olyan döntést hoznak, amelyet egyébként nem hoztak volna meg, és amely jelentős kárt okoz vagy valószínűsíthető módon okozhat annak a személynek, más személynek vagy csoportnak.
- Egy személy vagy egy adott csoport sebezhetőségének kihasználása koruk, fogyatékosságuk, szociális vagy gazdasági helyzetük miatt, azzal a céllal vagy eredménnyel, hogy jelentősen torzítsák a viselkedésüket, ami jelentős kárt okoz vagy okozhat annak a személynek vagy más személynek.
- Szociális pontozás MI-rendszerek használatával, amelyek a személyeket vagy csoportokat személyes jellemzőik, társadalmi viselkedésük és tevékenységeik alapján értékelik és jellemzik, ami a kezdeti kontextustól független vagy indokolatlan és aránytalan következtetésekhez vezethet.
- Prediktív bűnüldözés, amely természetes személyek kockázatának felmérésére vagy bűncselekmény elkövetésének előrejelzésére szolgáló profilozás személyiség és jellemzők felmérése alapján, kivéve, ha az ilyen MI-rendszert egy személy bűncselekményhez való kapcsolódása emberi értékelésének támogatására használják, amelynek igazolható kapcsolata van a bűncselekménnyel.
- Arcképekből álló adatbázisok létrehozásához vagy bővítéséhez az interneten elérhető arcképek vagy videómegfigyelési felvételek felhasználásával.
- Személy érzelmi állapotának felmérése, amely vonatkozik a munkahelyi vagy oktatási MI-rendszerekre, kivéve egészségügyi vagy biztonsági okokból, például annak észlelésére, hogy egy sofőr elalszik-e.
- Személyek biometrikus kategorizálása érzékeny jellemzők alapján, beleértve a nemi, faji, politikai irányultságot, vallást, nemi életet, szexuális irányultságot és filozófiai meggyőződéseket.
- Valós idejű távoli biometrikus azonosítás nyilvános helyeken, amely magában foglalja a biometrikus azonosító rendszerek teljes tilalmát, beleértve az utólagos azonosítást is, kivéve a bűnüldözést bírói jóváhagyással és a Bizottság felügyelete mellett előre meghatározott célokra, mint például bűncselekmény áldozatainak célzott keresése, terrorizmus megelőzése, súlyos bűncselekmények vagy gyanúsítottak célzott keresése, beleértve az emberkereskedelmet, szexuális kizsákmányolást, fegyveres rablást és környezet és természet elleni bűncselekményeket.

Kiberbiztonsági és adatvédelmi kötelezettségek a mesterségesintelligencia-rendszerekben

A Mesterséges intelligenciáról szóló jogszabály kötelezi a magas kockázatú MI-rendszerek szolgáltatóit és a rendszerszintű GPAI-szolgáltatókat a kiberbiztonsági kockázatok kezelésére egy mesterségesintelligencia-irányítási rendszer részeként.[7] A magas kockázatú MI-rendszereket a kiberbiztonsági követelményeknek megfelelően tanúsítani kell az (EU) 2019/881 rendelet szerint, a rendelet 15. cikkében meghatározott kiberbiztonsági követelmények alapján. A kapcsolódó szabványosítási kérelmet az Európai Szabványügyi Bizottság (European Committee for Standardization, CEN) és az Európai Elektrotechnikai Szabványügyi Bizottság (European Committee for Electrotechnical Standardization, CENELEC) kapta, azzal a követelménnyel, hogy a szabványosítás egyes területein konzultáljon az Európai Távközlési Szabványosítási Intézettel (European Telecommunications Standards Institute, ETSI).[8]

Másrészről a Mesterséges intelligenciáról szóló jogszabály kimondja, hogy a kritikus infrastruktúra üzemeltetésében részt vevő MI-rendszerek magas kockázatú kategóriába tartoznak. Ezáltal a rendelet egyértelműen megteremti a kapcsolatot a 2022/2557/EU irányelvvel,[9] amely a kritikus entitások rezilienciáját (*critical entities resilience,* CER) szabályozza hatálybalépését követően, felváltva az elavulttá vált korábbi irányelvet.[10] A kritikus entitások alapvető szolgáltatásokat nyújtanak a társadalmi funkciók fenntartásában, a gazdaság támogatásában, a közegészség és biztonság biztosításában, valamint a környezet megőrzésében.

Tekintettel arra, hogy a NIS2 irányelv[11] megteremti a kapcsolatot a CER irányelvvel, azaz a NIS2 irányelv I. és II. mellékletében meghatározott ágazatok és kapcsolódó alágazatok jelentős hányada kritikus infrastruktúraként jelenik meg, a kritikusinfrastruktúra-szolgáltatók kötelesek a NIS2 irányelvben meghatározott kiberbiztonsági követelményeket teljesíteni.

A NIS2 irányelv célja a kiberbiztonsági képességek fejlesztése az Európai Unió, a tagállamok és a hatályba tartozó vállalatok tekintetében, ezáltal a jogalkotó célja a jogszabállyal az egységes kiberbiztonsági szint növelése. Ennek értelmében e vállalatoknak meg kell felelniük a követelményeknek függetlenül attól, hogy használnak-e MI-rendszert vagy sem, illetve függetlenül az alkalmazott MI-rendszer kockázati szintjétől. A NIS2 hatálya alá tartozó szervezeteknek megfelelő és arányos intézkedéseket kell tenniük a hálózati és információs rendszereik biztonságát jellemző kockázatok kezelésére, valamint az incidensek megelőzésére és az incidensek hatásainak enyhítésére adminisztratív és kikényszerítő technológiai és fizikai kontrollok formájában. Ennek megfelelően ezeknek az intézkedéseknek a részét képezi a kockázatelemzés, az információs rendszerek biztonsági politikája és szabályzata, kockázatalapú biztonsági program kitűzése és megvalósítása, az incidenskezelés, üzletmenet-folytonosság, ellátási lánc biztonsága, kiberhigiéniai gyakorlatok

[7] Junklewitz et al. 2023; Soler Garrido et al. 2023.
[8] Junklewitz et al. 2023.
[9] Európai Parlament és a Tanács (EU) 2022/2557 irányelve.
[10] A Tanács 2008/114/EK irányelve.
[11] Európai Parlament és a Tanács 2022/2555 irányelve.

és kiberbiztonsági képzés, titkosítás, humánerőforrás-biztonság, hozzáférés-ellenőrzési politikák és eszközgazdálkodás, többfaktoros hitelesítés vagy folyamatos hitelesítési megoldások, biztonságos hang-, video- és szöveges kommunikációk, valamint biztonságos vészhelyzeti kommunikációs rendszerek használata. Emellett a vezetőknek elegendő ismeretekkel és készségekkel kell rendelkezniük ahhoz, hogy azonosítani tudják a szervezetükre vonatkozó kockázatokat, és értékelni tudják a kiberbiztonsági intézkedéseket és azok hatását a szervezetükre.

Továbbá, bár a NIS2 meghatározza a biztonsági szolgáltatókra vonatkozó követelményeket, a Cyber Solidarity Act javaslata kifejezetten előírja számukra a követelményeket, amelyek szerint az Európai Kiberpajzsban részt vevő entitásoknak korszerű és rendkívül biztonságos eszközökkel, felszerelésekkel és infrastruktúrákkal kell rendelkezniük. Ez lehetővé teszi a kollektív észlelési képességek, valamint a hatóságoknak és releváns entitásoknak szóló időben történő figyelmeztetések javítását, különösen a legújabb mesterséges intelligencia és adatelemzési technológiák használatával.[12]

Ahogy a NIS2 irányelv, úgy a Mesterséges intelligenciáról szóló jogszabály is meghatározza az adatvédelemhez való viszonyát is, hivatkozva a 2016/679/EU rendeletre,[13] azaz az Általános adatvédelmi rendeletre (GDPR). A GDPR az adatkezelési műveletekre vonatkozóan alapvető elveket, illetve az érintettek számára alapvető jogokat határoz meg. A személyes adatok kezelésére vonatkozó elveket az 5. cikk definiálja, amelyek:

- Jogszerűség, tisztességes eljárás és átláthatóság: a személyes adatokra vonatkozó adatkezelést jogszerűen, tisztességesen és az érintettek számára átlátható módon kell végezni.
- Célhoz kötöttség: a személyes adatok gyűjtése csak előre meghatározott, egyértelmű és jogszerű célból történjen.
- Adattakarékosság: az adatkezelési tevékenység csak és kizárólag a szükséges személyes adatokra korlátozódjon.
- Pontosság: biztosítani kell az adatkezelési tevékenységben érintett személyes adatok naprakészségét.
- Korlátozott tárolhatóság: a személyes adatokat csak az adatkezelés meghatározott céljainak megfelelő ideig szabad tárolni.
- Integritás és bizalmas jelleg: a személyes adatok kezelése során megfelelő technikai vagy szervezési intézkedések alkalmazásával biztosítani kell a személyes adatok bizalmasságát, sértetlenséget és rendelkezésre állását.

[12] Európai Bizottság 2023.
[13] Európai Parlament és a Tanács 2016/679 rendelete.

Az adatkezelő és az adatfeldolgozó tevékenysége során végzett adatkezelési műveletek a 6. cikk értelmében akkor tekinthetők jogszerűnek, ha:

- az érintett hozzájárulását adta;
- az adatkezelés olyan szerződés teljesítéséhez szükséges, amelyben az érintett az egyik fél;
- az adatkezelés az adatkezelőre vonatkozó jogi kötelezettség teljesítéséhez szükséges;
- az adatkezelés az érintett vagy egy másik természetes személy létfontosságú érdekeinek védelme miatt szükséges;
- az adatkezelés közérdekű vagy az adatkezelőre ruházott közhatalmi jogosítvány gyakorlásának keretében végzett feladat végrehajtásához szükséges, vagy
- az adatkezelés az adatkezelő vagy egy harmadik fél jogos érdekeinek érvényesítéséhez szükséges, kivéve, ha ezen érdekekkel szemben elsőbbséget élveznek az érintett olyan érdekei vagy alapvető jogai és szabadságai, amelyek személyes adatok védelmét teszik szükségessé, különösen, ha az érintett gyermek.

Irányítási rendszerek

A Mesterséges intelligenciáról szóló jogszabály megköveteli a magas kockázatú MI-rendszerek szolgáltatóitól, hogy hozzanak létre és kezeljenek egy MI-irányítási rendszert (*artificial intelligence management system,* AIMS), amely a minőségirányítással, az információbiztonsággal vagy kiberbiztonsággal, valamint a személyes adatokra vonatkozó adatkezeléssel és adatfeldolgozással is együtt dolgozik. Ezek a követelmények nyilvánvaló hasonlóságot mutatnak az ISO/IEC 42001 szabványban meghatározottakkal.

Az ISO/IEC 42001[14] az AIMS követelményeit határozza meg, azaz útmutatást nyújt egy ilyen irányítási rendszer létrehozásához, megvalósításához, fenntartásához és folyamatos fejlesztéséhez. Átfogó keretet biztosít az MI-rendszerek etikus kialakításához, és biztosítja, hogy az MI-technológiák megfeleljenek az átláthatóság, elszámoltathatóság és adatvédelem elveinek. Az AIMS együttműködik a minőségirányítási rendszerrel (*quality management system,* QMS), az információbiztonsági irányítási rendszerrel (*information security management system,* ISMS) és a személyes adatkezelési rendszerekkel (*privacy information management system,* PIMS). Ennek megfelelően az AIMS integrálódik a szervezeti folyamatokba, illetve több ISO szabvánnyal, úgymint az ISO/IEC 27001[15] az ISMS, az ISO 9001[16] a QMS és az ISO/IEC 27701[17] a PIMS tekintetében. (Az ISO/IEC 27701 az ISO/IEC 27001 kiterjesztése, amelyet a jövőben a független ISO/IEC DIS 27701 vált fel.) Az ilyen integrációs képességek nem meglepők, mivel az ISO 42001 nagy hangsúlyt fektet arra, hogy az MI-rendszerek megfeleljenek a jogszabályi elvárásoknak, közte az adatvédelmi

[14] ISO/IEC 42001:2023.
[15] ISO/IEC 27001:2022.
[16] ISO 9001:2015.
[17] ISO/IEC 27701:2019.

követelményeknek, valamint kiberbiztonsági intézkedések végrehajtását követeli meg az MI-rendszerek fenyegetettségek elleni védelme érdekében.

Irányítási rendszer lévén az AIMS alapvető eleme a kockázatmenedzsment, amely rendszerezett megközelítést biztosít az MI életciklusa során felmerülő kockázatok azonosítására, elemzésére és mérséklésére, felhasználva az MI hatásvizsgálat-eredményeit. A szabvány technikai útmutatást nyújt az irányítási rendszer szervezeti célokból és etikai normákból történő származtatásához, beleértve az MI-rendszerek folyamatos monitorozására és fejlesztésére vonatkozó eljárásokat. Ez biztosítja, hogy az etikai megfontolások integrálódjanak az irányítási rendszerbe, beleértve az MI-megoldás fejlesztését és használatát szabályozó etikai irányelveket és egy felügyeleti mechanizmust. Mivel egy MI-rendszer tanítása torzított adatkészlet miatt helytelen módon is végbemehet, az AIMS alapja a különböző és reprezentatív adatkészlet használata az MI-algoritmusok működése torzításának csökkentése érdekében. Végül az átláthatósági és elszámoltathatósági követelmények megkövetelik az MI-algoritmusok, adatforrások és döntéshozatali folyamatok dokumentálását.

A mesterséges intelligencia kockázatai

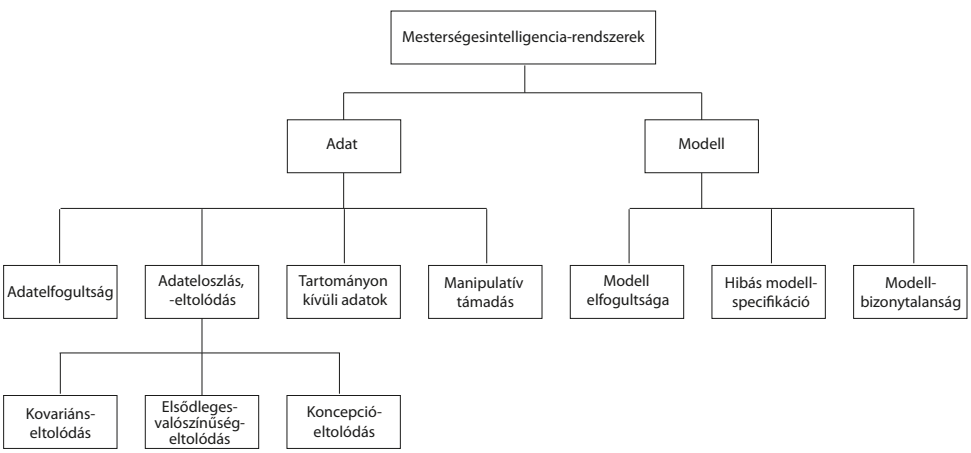
A NIST MI Kockázatkezelési Keretrendszere (*artificial intelligence risk management framework,* AI RMF) önkéntes használatra készült, és célja, hogy javítsa a megbízhatósági szempontok beépítésének képességét az MI-termékek, szolgáltatások és rendszerek tervezésébe, fejlesztésébe, használatába és értékelésébe. Az irányítás kritikus szerepet játszik az MI-kockázatkezelés minden más szakaszában, egy olyan kultúra kialakításával, amely felismeri a mesterséges intelligenciával kapcsolatos potenciális kockázatokat. Az irányítás lépése magában foglalja a kockázatok kezelésére és hatásuk felmérésére szolgáló folyamatok és dokumentációk kidolgozását és megvalósítását.[18]

A kockázatmenedzsment ennélfogva magában foglalja az átfogó MI-kockázatelemzést a felhasználókra, a társadalomra és még a környezeti hatásokra vonatkozó lehetséges hatások azonosítása érdekében, valamint az azonosított kockázatok mérséklése és a negatív hatások minimalizálása érdekében az intézkedések kidolgozását és végrehajtását. Az MI-kockázatértékelésnek tartalmaznia kell az MI-rendszert használó vállalatnak a rendszerre vonatkozó kiberbiztonsági és adatvédelmi kockázatkezelését, valamint a teljes vállalati kockázatkezelés részét kell képeznie. Mint általában, a kapcsolódó kockázatokat a fenyegetés valószínűsége és a hatás súlyossága alapján számoljuk, alkalmazva a Mesterséges intelligenciáról szóló jogszabály 3. cikk (2) bekezdésében foglaltakat, a kockázat az ártalom bekövetkezési valószínűségének és az ártalom súlyosságának kombinációja.[19]

[18] Tabassi 2023.
[19] Európai Parlament 2024.

Az MI-rendszereket célzó, illetve általa jelentett kockázatokat technikai, operatív, etikai és szabályozási kockázatok széles körébe lehet sorolni. Technikai kockázatot jelent például egy MI-rendszer bemeneteinek manipulálása a rendszer megtévesztésére, az algoritmusokban található hibák vagy gyengeségek, a modell kimenetének korrupttá tétele adatmérgezés útján. Operatív kockázatok közé tartozik egy MI-rendszer telepítése, frissítése és karbantartása során fellépő hibák, az adathalmazok vagy összetettebb feladatok skálázására vonatkozó hiányosságok, illetve a meglévő technológiák vagy folyamatok közötti inkompatibilitás. Etikai kockázatot jelent például a rendszer tanításához felhasznált adatokban meglévő elfogultság, részrehajlás továbbvitele vagy felerősítése, a döntéshozatali folyamatok megértésének vagy magyarázatának képtelensége (feketedoboz-probléma), az MI által vezérelt döntések és cselekvések felelősségének megállapításával kapcsolatos problémák, valamint összefüggésben az adatvédelmi kérdésekkel, az egyének magánéletének megsértése túlzott adatgyűjtés és megfigyelés révén. Szabályozási (vagy jogszabályi megfelelés) kockázatok közé sorolandó a jogszabályi előírások és szabványok követelményeitől való eltérés, azok ellenében történő nem megfelelés, az MI-rendszerek által okozott károk miatt indított jogi eljárások és bírságok.[20]

Zhang és munkatársai[21] az MI-rendszerek kapcsán felmerülő kockázatokat forrásuk szerint két nagy csoportra osztották (2. ábra): az adat és a modell. Az egyes kockázatok leírását az 1. táblázat ismerteti.



2. ábra: Mesterséges intelligencia kapcsán felmerülő kockázatok osztályozása
Forrás: a szerző szerkesztése Zhang et al. 2022 alapján

[20] Európai Parlament 2024.
[21] Zhang et al. 2022.

1. táblázat: Mesterséges intelligencia kapcsán felmerülő adat- és modellkockázatok

Kockázat	**Leírás**
Adatelfogultság	Az adatelfogultság arra utal, amikor az MI-modellekben egyes csoportok vagy elemtípusok túl- vagy alulreprezentáltak. A túlreprezentált jelleg esetén az adott csoport- vagy elemtípus nagyobb súlyt kap, mint mások. Ha egy adott osztály vagy csoport alulreprezentált, a modell gyengén teljesíthet ezen csoport kapcsán. Tekintettel arra, hogy az MI-modellek általában múltbeli adatok alapján tanulnak döntéseket hozni, gyakran az adott adatokban meglévő elfogultságokat továbbörökítik.
Adateloszlás, -eltolódás	Az adateloszlás, -eltolódás azt a helyzetet jellemzi, amikor az MI-modell tanítási és futásidőbeli adatai eltérő eloszlásokat mutatnak: • Kovarianciaeltolódás esetén a címkeelosztások eltérők, de a címkék jellemzői megegyeznek. • Elsődleges valószínűség eltolódása esetén a jellemzők eloszlása eltérő, de a jellemzőkhöz adott címkék azonosak. • Koncepcióeltolódás esetén a jellemzők ugyanazok, de a jellemzőkkel rendelkező címkék eltérők.
Tartományon kívüli adatok	Az inputadatokra vonatkozó nem megfelelő validálás és kezelés esetén nagy a valószínűsége annak, hogy a betanított modell téves előrejelzéseket ad magas bizonyossággal.
Manipulatív támadás	A manipulatív támadás célzott és nem célzott támadást foglal magában. A célzott támadás célja, hogy az MI-modell egy olyan ellenkező képet osztályozzon be célzott osztályként, amelynek valódi címkéje eltér ettől, ezt szándékos tervezéssel érik el (például adatmanipulációval). A nem célzott támadás célja, hogy az MI-modell olyan előrejelzést adjon, amely eltér a valódi címkétől anélkül, hogy meghatározott célt állítanának fel.
Modell elfogultsága	A tanítási adatokra jellemző elfogultság vagy hibás algoritmus következtében a betanított MI-modell elfogulttá válik.
Hibás specifikáció	A modell rossz specifikációja akkor következik be, ha a modell feltételezései nem megfelelők a tanítás során használt adatokhoz. A modell rossz specifikációját három tényező okozhatja: • Modellformalizálási hiba: a megadott funkcionális forma nem megfelelő a valódi kapcsolat jellemzéséhez, azaz bár minden magyarázó változó rendelkezésre áll, de a modell nem képes helyesen jellemezni a magyarázó változók és a magyarázott változó közötti kapcsolatot. • Modell túltanulása: a feladathoz mérten a szükségesnél összetettebb modell illesztése miatt a modell, bár kiválóan teljesít a tanítási adatok illesztésében, a tanítási adatokon túli teljesítménye az elvártakhoz képest alulmarad. • Változóbevonási hiba: változóbevonási hiba merül fel, amikor egy szükséges változó a modellből kimarad, vagy egy nem szükséges változó szerepel a modellben (tévedésből vagy szándékosan). A kimaradt jelentős változó eredményeként a modell nem képes megfelelően jellemezni az adatokat, ami végül elfogultságot eredményez. Az irreleváns változó bevonása a modell túltanulásához vezethet.
Modellbizonytalanság	Minden előrejelzési modell definíció szerint a valóság idealizált reprezentációja, és ezért eredendően nem képes tökéletesen reprezentálni a valódi rendszer viselkedését. A modell előrejelzési bizonytalansága a modell paramétereiben és szerkezetében rejlő sajátosság.

Forrás: a szerző szerkesztése Zhang et al. 2022 alapján

A MITRE ATLAS (MITRE 2024) hatás (impact) taktika részét képező technikák számos módszert részletesen tárgyalnak (2. táblázat). Az egyes technikák a modell, illetve az MI-rendszer hibás működéséből fakadó addicionális hatásokat („Külső hatás") fedik le. Az adatokra és a modellt megvalósító algoritmusra vonatkozó bizalmasság, sértetlenség és rendelkezésre állás kérdésével a MITRE ATLAS további taktikái foglalkoznak (például „ML Model Access", azaz „Gépi tanulás modell hozzáférés").

2. táblázat: MITRE ATLAS hatás- (impact) technikák

Technika	Leírás
Gépi tanulási modell kijátszása	A támadó megakadályozza, hogy a gépi tanulási modell helyesen azonosítsa az adatok tartalmát. Ez a technika felhasználható a gépi tanulást alkalmazó feladatok kijátszására.
Gépi tanulás megtagadása	A támadó szándékosan olyan bemeneteket hoz létre, amelyek nagy mennyiségű haszontalan számítást igényelnek a gépi tanulási rendszertől, ezzel lerontva vagy leállítva a szolgáltatást.
Gépi tanulás modell integritásának erodálása	A támadó manipulált adatokkal lerontja a modell teljesítményét, ezzel idővel megingatva a rendszerbe vetett bizalmat.
Költségnövelés	A támadó különböző gépi tanulási szolgáltatásokat célozhat meg haszontalan kérdésekkel vagy számításigényes bemenetekkel, hogy növelje a szolgáltatások futtatásának költségeit az áldozatszervezet számára.
Gépi tanulási rendszer hamis adatokkal történő bombázása	A támadó a gépi tanulási rendszert hamis adatokkal arra kényszerítheti a rendszert használó felett, hogy helytelen következtetések felülvizsgálatára és kijavítására fordítsa erőforrásait.
Külső károk	A támadó egy MI-rendszer erőforrásait vagy képességeit felhasználhatja saját céljának elérésére, miközben külső károkat okoz. Ezek a károk érinthetik a szervezetet (például pénzügyi károk, hírnévrombolás), annak felhasználóit (például felhasználói károk) vagy egy tágabb közösséget (például társadalmi károk). A pénzügyi kár magában foglalhatja a vagyon, tulajdon vagy egyéb pénzügyi eszközök elvesztését lopás, csalás vagy hamisítás miatt, vagy a nyomásgyakorlást, hogy pénzügyi erőforrásokat biztosítsanak a támadó számára. A hírnévrombolás a közvélemény és a szervezet iránti bizalom csökkenését jelenti. A társadalmi károk olyan negatív hatásokat eredményezhetnek, amelyek a közvéleményt vagy specifikus sebezhető csoportokat érinthetnek, mint például a gyermekek káros tartalommal való találkozása. A felhasználói károk magukban foglalhatnak pénzügyi és hírnévkárosodást, amelyeket az egyéni áldozatok éreznek, nem pedig szervezeti szinten jelentkeznek. A támadó az MI-rendszerek modelljének és a bemeneti, illetve kimeneti adatok által jelentett szellemi tulajdont lophatnak, gazdasági kárt okozva az áldozat szervezetnek.

Forrás: a szerző szerkesztése MITRE 2024 alapján

Kutatási eredmények

Ennek értelmében további kategóriaként jelennek meg az adatvédelmi és a kiberbiztonsági kockázatok. Az adatvédelemi kockázatok alatt többek közt a személyes adatok jogosulatlan felhasználása, nem megfelelő intézkedések az adatvédelem biztosítására, valamint a nem megfelelő jogalap alkalmazása értendő. Kiberbiztonsági kockázatként

jelenik meg az adott információs és kommunikációs technológia (IKT) vonatkozásában, például a jogosulatlan hozzáférés, az adatszivárgás, illetve a szolgáltatásmegtagadás-jellegű támadás formájában. A fenyegető tényezők mindehhez kihasználhatják a fizikai környezet, az emberek (alkalmazottak és harmadik felek), a technológia vagy akár a folyamatok gyengeségeit. A kiberbiztonsági területen végzett tevékenységek közvetlen hatást gyakorolhatnak egy IKT-rendszer integritására vagy rendelkezésre állására, ugyanakkor közvetlenül befolyásolhatják az adatok bizalmasságát, valamint az IKT-rendszerek integritását és rendelkezésre állását. Az IKT-rendszer szintjén felmerülő problémák negatív hatással lehetnek az adatokra, és ilyen problémák könnyen érinthetik az MI-szintet az MI-algoritmus integritása vagy a tanulási adatok mérgezése révén. A költség begyűjtése ezen a szinten jelentős közvetlen hatást gyakorol. A legfelső szinten, amely az üzleti hatásokat képviseli, mindezek pénzügyi hatásokkal járhatnak. A jogszabályi követelmények e kockázatok tükrében szükségesek és a kockázati profilokhoz rendelt elvárások miatt arányosak is.

Mindezek értelmében az MI-rendszerek minőségi kockázatai szerteágazók (3. táblázat), befolyásolják az adatok integritását, a modellek pontosságát, a folyamatok megbízhatóságát, az eredmények megbízhatóságát és a felhasználói interakciót.

3. táblázat: Az MI-rendszerek minőségi kockázatai

Kockázat	**Leírás**
Adatok minőségi kockázatai	• Adatelfogultság: a valós élethelyzeteket nem tükröző adathalmaz elfogult MI-modellekhez vezethet. • Adatinkonzisztencia: az inkonzisztens adatformázás, címkézés vagy kategorizálás hibákat okozhat az MI-modell tanításában. • Adathiányosság: a hiányos vagy nem teljes adathalmaz olyan MI-modellekhez vezethet, amelyek nem képesek hatékonyan kezelni bizonyos helyzeteket. • Adatpontosság: a pontatlan adatok helytelen döntésekhez vezethetnek.
Modell minőségi kockázatai	• Modell túltanulása: az MI-modellek, amelyek túl szorosan illeszkednek a tréningadatokhoz, nem általánosíthatók jól az új, ismeretlen adatokra. • Modell alultanulása: az egyszerűsített modellek, amelyek nem képesek megragadni az adatok mögöttes mintázatait, gyenge teljesítményt eredményezhetnek. • Algoritmus elfogultsága: az algoritmusokban lévő hibák elfogult, hibás döntéshozatalt eredményezhetnek. • Robusztusság hiánya: az MI-modellek, amelyek nem robusztusak, nem képesek ellenállni az enyhe változásoknak vagy támadásoknak. • Magyarázhatóság és értelmezhetőség: a modellek döntéshozatali folyamatai megértésének nehézsége (feketedoboz-jelenség) csökkentheti a bizalmat és a felelősséget.

Kockázat	Leírás
Folyamat minőségi kockázatai	• Rossz tréningfolyamatok: az elégtelen tanítási eljárások szuboptimális-modell-teljesítményhez vezethetnek. • Elégtelen validáció és tesztelés: a szigorú validáció és tesztelés hiánya hibás modellekhez vezethet. • Elégtelen monitorozás: az MI-rendszerek folyamatos monitorozásának elmulasztása idővel észrevétlen minőségi romláshoz vezethet. • Változáskezelési problémák: Az MI-rendszerek frissítésének és változtatásainak kezelési folyamatai hiányosak lehetnek, ami a minőség degradálásához vezethet.
Eredmény minőségi kockázatai	• Teljesítményromlás: idővel az MI-modellek teljesítménye romolhat, ha nem megfelelő a karbantartás és a frissítés. • Nem szándékolt következmények: az MI-rendszerek olyan következtetést tehetnek, amelyek technikailag helyesek, de kontextuálisan nem megfelelők vagy károsak. • Megbízhatósági problémák: azok az MI-rendszerek, amelyek nem teljesítenek következetesen, aláásható a bizalmat és a megbízhatóságot. • Méretezhetőségi problémák: a hatékony skálázódásra nem képes MI-rendszerek nehézségekbe ütközhetnek a növekvő adatmennyiségek vagy felhasználók kezelésében.
Felhasználói interakció kockázatai	• Felhasználói félreértés: a felhasználók félreérthetik a kimenetet, ami helytelen döntésekhez vagy cselekvésekhez vezethet. • Felhasználói bizalom hiánya: ha a felhasználók nem bíznak az MI-rendszerben, vonakodhatnak használni azt, csökkentve annak hatékonyságát. • Használhatósági problémák: rosszul tervezett interfészek és interakciók akadályozhatják az MI-rendszerek hatékony használatát.

Forrás: a szerző szerkesztése

Összegzés és konklúzió

A tanulmány betekintést nyújtott a mesterségesintelligencia- (MI-) rendszerek megvalósításának az Európai Unióban körvonalazódó jogi keretébe. A jogszabályi környezet átfogó megközelítést alkalmaz, amely definiálja a különböző kockázati profilokat, amelyekhez a kockázati szint függvényében kötelezettségeket ír elő.

A Mesterséges intelligenciáról szóló jogszabály megköveteli a magas kockázatú MI-rendszerek szolgáltatóitól egy MI-irányítási rendszer (AIMS) létrehozását és működtetését, amely integrálja a minőségirányítást, az információbiztonságot és az adatvédelmet. A jogszabályi kötelezettségek felépítése és jellege és az ISO/IEC 42001 szabvány között jelentős átfedés azonosítható, így a szabvány útmutatást nyújt az AIMS létrehozásához, fenntartásához és fejlesztéséhez. Egy AIMS célja, hogy biztosítsa egy MI-rendszer etikus, átlátható működését, amely megfelel az adatvédelmi törvényeknek, integrálódva más irányítási rendszerekhez, mint az ISO/IEC 27001 (ISMS), ISO 9001 (QMS) és ISO/IEC 27701 (PIMS).

Mint minden irányítási rendszer, így az AIMS egyik alapvető eleme a kockázatmenedzsment, amely az MI életciklusa során felmerülő kockázatok kezelésére és mérséklésére szolgál. A NIST MI Kockázatkezelési Keretrendszere (AI RMF) önkéntes

használatra készült, és célja a megbízhatóság növelése az MI-rendszerek tervezése, fejlesztése és használata során. A kockázatmenedzsment magában foglalja a kockázatok azonosítását, elemzését és a mérséklő intézkedések kidolgozását, figyelembe véve a technikai, operatív, etikai és szabályozási kockázatokat. A kockázatok közé tartoznak például a bemeneti adatok manipulálása, algoritmikus hibák, adatelfogultság és a döntéshozatali folyamatok átláthatatlansága.

Az MI-rendszerek minőségi kockázatai szerteágazók, befolyásolják az adatok integritását, a modellek pontosságát, a folyamatok megbízhatóságát, az eredmények megbízhatóságát és a felhasználói interakciót. Ezek közé tartozik az adatelfogultság, adatinkonzisztencia, adathiányosság, modell túltanulása vagy alultanulása, algoritmikus elfogultság és az MI-rendszerek folyamatos monitorozásának elmulasztása. Az eredmények minőségi kockázatai között szerepel a teljesítményromlás, nem szándékolt következmények és megbízhatósági problémák. A felhasználói interakció kockázatai közé tartozik a felhasználói félreértés, bizalom hiánya és használhatósági problémák.

Felhasznált irodalom

Council of the European Union (2024): *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Analysis of the Final Compromise Text with a View to Agreement.* Online: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

Európai Bizottság (2023): *Javaslat. Az Európai Parlament és a Tanács rendelete a kiberbiztonsági fenyegetések és események észlelése, valamint az azokra való felkészülés és reagálás céljából az Unión belüli szolidaritás és képességek megerősítését célzó intézkedések meghatározásáról*. Online: https://eur-lex.europa.eu/legal-content/HU/TXT/HTML/?uri=CELEX:52023PC0209

Európai Parlament (2024): *HELYESBÍTÉS az Európai Parlament által 2024. március 13-án a mesterséges intelligenciára vonatkozó harmonizált szabályok megállapításáról.* Online: www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_HU.pdf

Európai Parlament és a Tanács 2016/679 rendelete a természetes személyeknek a személyes adatok kezelése tekintetében történő védelméről és az ilyen adatok szabad áramlásáról. Az Európai Parlament és a Tanács (EU) 2016/679 rendelete.

Európai Parlament és a Tanács 2019/881 rendelete az ENISA-ról (az Európai Uniós Kiberbiztonsági Ügynökségről) és az információs és kommunikációs technológiák kiberbiztonsági tanúsításáról. Az Európai Parlament és a Tanács (EU) 2019/881 rendelete.

Európai Parlament és a Tanács 2022/2481 határozata a Digitális évtized 2030 szakpolitikai program létrehozásáról.

Európai Parlament és a Tanács 2022/2555 irányelve az Unió egész területén egységesen magas szintű kiberbiztonságot biztosító intézkedésekről.

Európai Parlament és a Tanács (EU) 2019/790 irányelve (2019. április 17.) a digitális egységes piacon a szerzői és szomszédos jogokról, valamint a 96/9/EK és a 2001/29/EK irányelv módosításáról.

Európai Parlament és a Tanács (EU) 2022/2557 irányelve (2022. december 14.) a kritikus szervezetek rezilienciájáról.

European Commission (2024): *AI Act.* Online: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

ISO 9001:2015 Quality Management Systems — Requirements.

ISO/IEC 27001:2022 Information Security, Cybersecurity and Privacy Protection. Information Security Management Systems. Requirements.

ISO/IEC 27701:2019 Security Techniques Extension to ISO/IEC 27001 and ISO/IEC 27002 for Privacy Information Management Requirements and Guidelines.

ISO/IEC 42001:2023 Information Technology Artificial Intelligence Management System.

ISO/IEC DIS 27701 Information Security, Cybersecurity and Privacy Protection Privacy Information Management Systems Requirements and Guidance.

Jiang, Yuchen et al. (2022): Quo Vadis Artificial Intelligence? *Discover Artificial Intelligence,* 2(4). Online: https://doi.org/10.1007/s44163-022-00022-8

Junklewitz, Henrik et al. (2023): *Cybersecurity of Artificial Intelligence in the AI Act*. Luxembourg: Office of the European Union. Online: https://doi.org/10.2760/271009

MITRE (2024): MITRE ATLAS. Online: https://atlas.mitre.org/

Soler Garrido, Josep et al. (2023): *Analysis of the Preliminary AI Standardisation Work Plan in Support of the AI Act*. Luxembourg: Office of the European Union. Online: https://doi.org/10.2760/5847

Tabassi, Elham (2023): *Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST Trustworthy and Responsible AI, National Institute of Standards and Technology.* Gaithersburg. Online: https://doi.org/10.6028/NIST.AI.100-1

Tanács 2008/114/EK irányelve az európai kritikus infrastruktúrák azonosításáról és kijelöléséről, valamint védelmük javítása szükségességének értékeléséről. 2008/114/EK (2008).

Zhang, Xiaoge et al. (2022): Towards Risk-Aware Artificial Intelligence and Machine Learning Systems: An Overview. *Decision Support Systems, 159*, 113800. Online: https://doi.org/10.1016/j.dss.2022.113800