Journal of the Association for Information Systems (2021) **22**(2), **325-352**.
doi: 10.17705/1jais.00664

KUTATÁSI CIKK

ISSN 1536-9323

A mesterséges intelligencia szociotechnikai kibontakozása: Megközelítés a kiismerhetetlen mesterséges intelligencia rendszerek szervezeti bevezetéséhez

Aleksandre Asatiani1, Pekka Malo2, Per Rådberg Nagbøl3, Esko Penttinen4, Tapani Rinta-Kahila5, Antti Salovaara6

[1]University of Gothenburg, Svédország, aleksandre.asatiani@ait.gu.se[2] Aalto University School of Business, Finnország, pekka.malo@aalto.fi[3] IT University of Copenhagen, Dánia, pena@itu.dk
[4]Aalto University School of Business, Finnország, esko.penttinen@aalto.fi
[5]The University of Queensland, Ausztrália, t.rintakahila@uq.edu.au
[6]Aalto University School of Arts, Design and Architecture, Finnország, antti.salovaara@aalto.fi

Absztrakt

A tanulmány egy olyan megközelítést mutat be, amely lehetővé teszi a megfejthetetlen (azaz megmagyarázhatatlan) mesterséges intelligencia (AI), például a neurális hálózatok felelős és biztonságos megvalósítását a szervezeti környezetben. Egy feltáró esettanulmányra és a nemrégiben javasolt burkolás fogalmára támaszkodva leír egy olyan esetet, amikor egy szervezet sikeresen "burkolta" a mesterséges intelligencia megoldásait, hogy egyensúlyt teremtsen a rugalmas mesterséges intelligencia modellek teljesítménybeli előnyei és a kifürkészhetetlen modellekből adódó kockázatok között. A szerzők számos burkoló módszert mutatnak be - olyan világos határok kijelölése, amelyeken belül a mesterséges intelligenciának kölcsönhatásba kell lépnie a környezetével, a képzési adatok megfelelő kiválasztása és gondozása, valamint a bemeneti és kimeneti források megfelelő kezelése -, valamint ezek hatását a szervezeten belüli mesterséges intelligencia-modellek kiválasztására. Ez a munka két kulcsfontosságú hozzájárulást tesz: Bevezeti a szociotechnikai burkolás fogalmát, bemutatva, hogy egy szervezet sikeres mesterséges intelligencia burkolása milyen módon függ a társadalmi és technikai tényezők kölcsönhatásától, és ezzel a szakirodalom fókuszát a puszta technikai kérdéseken túlra terjeszti. Másodszor, az empirikus példák azt mutatják be, hogy a szociotechnikai burkolás operacionalizálása hogyan teszi lehetővé egy szervezet számára, hogy kezelje az alacsony megmagyarázhatóság és a magas teljesítmény közötti kompromisszumot, amelyet a kifürkészhetetlen modellek jelentenek. Ezek a hozzájárulások megnyitják az utat a felelősségteljesebb, elszámoltathatóbb mesterséges intelligencia szervezetekbeli megvalósításai előtt, amelyek révén az emberek jobban ellenőrizhetik még a kifürkészhetetlen gépi tanulási modelleket is.

Kulcsszavak: Mesterséges intelligencia, megmagyarázható mesterséges intelligencia, XAI, borítékolás, szociotechnikai rendszerek, gépi tanulás, közszféra.

Hind Benbya volt az elfogadó főszerkesztő. Ezt a kutatási cikket 2020. február 29-én nyújtották be, és három átdolgozáson esett át.

1 Bevezetés

A nagyméretű adatok és a gépi tanulási (ML) technológia fejlődése olyan mesterséges intelligenciát (AI) alkalmazó rendszereket eredményezett, amelyek jelentős hatékonyságnövekedést és újszerű információfeldolgozási képességeket biztosítanak az érintett szervezetek számára. Míg az ML-modellek képesek lehetnek felülmúlják az emberi szakértők teljesítményét igényes elemzési és döntési helyzetekben (McKinney et al., 2020), működési logikájuk drámaian különbözik az emberek hasonló problémákhoz való hozzáállásának módjától. Az adatmennyiség és a rendelkezésre álló számítási teljesítmény gyors növekedése egyre összetettebbé tette a mesterséges intelligencia rendszereket, ami a viselkedésüket kifürkészhetetlenné teszi, és ezért az ember számára nehezen értelmezhetővé és megmagyarázhatóvá teszi őket

(Faraj et al., 2018; Stone et al., 2016). Míg az ilyen rendszerek gazdasági értéke ritkán kétséges, a szélesebb körű szervezeti és társadalmi következmények, köztük az olyan negatív mellékhatások, mint a fel nem fedezett elfogultságok, kezdenek aggodalomra okot adni (Benbya et al., 2020; Brynjolfsson & McAfee, 2014; Newell & Marabelli, 2015). Így az emberek azon képessége, hogy meg tudják magyarázni, hogyan hozzák létre a mesterséges intelligencia rendszerek a kimeneteiket, amelyet "magyarázhatóságnak" neveznek (pl. Rosenfeld & Richardson, 2016), kiemelt kérdéssé vált a különböző területeken.

A mesterséges intelligencia rendszerek kifürkészhetetlensége számos etikai, jogi és gyakorlati kérdést vet fel. Az ML-modellek szükségszerűen esztelenül működnek, ami azt jelenti, hogy egyetlen nézőpontból közelítik meg a munkát, a tágabb kontextus tudatos megértése nélkül (Burrell, 2016; Salovaara et al., 2019). Az ML-modellek például nem tudnak reflektálni tevékenységük etikájára vagy jogszerűségére. Ennek megfelelően egy mesterséges intelligencia rendszer nem szándékolt elfogultságot és diszkriminációt mutathat, miután megtanult nem megfelelő tényezőket figyelembe venni a döntéshozatal során (Martin, 2019). Az ilyen problémák révén a képzési szakaszban és azon túl egy szervezet (akarva vagy akaratlanul) végül olyan módon működhet, amely ellentétes az értékeivel (Firth, 2019), a modellek pedig hajlamosak lehetnek a bosszantó etikai kérdésekkel - például bizonyos embercsoportok diszkriminációjával - kapcsolatos elfogultságokra és hibákra. A modellek szilárd etikai szemléletű tervezése eszközöket biztosíthatna az ilyen elfogultságok és hibák azonosítására, megítélésére és kijavítására (Martin, 2019), de mindez lehetetlen, ha a modell cselekedetei kifürkészhetetlenek. Az etikai kérdések mellett vannak olyan jogalkotási tényezők, amelyek konkrét és megkerülhetetlen követelményeket támasztanak a megmagyarázhatósággal szemben (Desai & Kroll, 2017). A hatóságoknak gyakran tiszteletben kell tartaniuk az átláthatóság követelményeit a tevékenységükkel kapcsolatban, és a magánvállalkozások is kötelesek lehetnek magyarázatot adni és indokolni például azt, hogy hogyan használják fel az ügyfelek adatait. Az Európai Unió általános adatvédelmi rendelete (GDPR) kiemelkedő példaként szolgál a közelmúltbeli jogalkotási intézkedésekre, amelyek elősegítik az érintettek azon jogát, hogy magyarázatot kapjanak a róluk gyűjtött adatokon alapuló bármely döntésről (Európai Unió, 2016).

Egy megmagyarázható mesterséges intelligencia rendszer létrehozása azonban nem mindig kivitelezhető. A megfejthetetlenségnek számos formája van, és olyan elemekhez kapcsolódik, mint a szándékos vállalati vagy állami titkolózás, a technikai analfabétizmus és az ML-modellek veleszületett jellemzői (Burrell, 2016). Ez a sokrétű jelleg, az emberi logika korlátaival együtt azt jelenti, hogy a megmagyarázhatósági problémákra nincsenek egyszerű megoldások (Edwards, 2018; Robbins, 2020). Egyes jogtudósok például azt állítják, hogy a GDPR magyarázathoz való jogra vonatkozó rendelkezése nem elegendő, és értelmetlen "átláthatóságot" eredményezhet, amely valójában nem felel meg a felhasználói igényeknek (Edwards & Veale, 2017): bár technikailag lehet magyarázat egy adott döntésre, ez nem biztos, hogy érthető az érintett személy(ek) számára. Bár a

az olyan megközelítések, mint a jogi auditálás (O'Neil, 2016; Pasquale, 2015), a robusztus rendszertervezés (Rosenfeld & Richardson, 2019) és a felhasználói oktatás bizonyos esetekben javíthatják a megmagyarázhatóságot, egydimenziósak és alkalmatlanok a mesterséges intelligencia esztelen működése által jelentett alapvető kihívások kezelésére (Burrell, 2016). Szervezeti környezetben az informatikai (IT) rendszerek az érdekeltek széles körét érintik, akik eltérő, gyakran élesen ellentétes igényeket és elvárásokat mutatnak (Koutsikouri et al., 2018). Az AI-ügynökök viselkedésének magyarázatát tovább bonyolítja az a környezet, amelyben az AI-fejlesztés zajlik, a különböző meglévő munkafolyamatokkal, struktúrákkal, hierarchiákkal és örökölt technológiákkal. Ezek a kihívások a magyarázhatóság emberközpontú és pragmatikus megközelítései iránti felhívást váltottak ki (Mittelstadt et al., 2019; Ribera & Lapedriza, 2019). Ez arra hívja fel a figyelmünket, hogy a megmagyarázhatóságot szociotechnikai perspektívából közelítsük meg, hogy figyelembe vegyük a technológia, az emberek, a folyamatok és a szervezeti elrendezések összekapcsolódó természetét, és ezáltal kiegyensúlyozott figyelmet fordítsunk a technológia instrumentális és humanista eredményeinek egyaránt (Sarker et al., 2019).

Ennek fényében a következő kutatási kérdésre keressük a választ: *Hogyan tudja egy szervezet biztonságosan és társadalmilag felelősségteljesen kihasználni a kifürkészhetetlen mesterséges intelligencia rendszereket?* Vizsgálatunkat az a vágy inspirálta, hogy megértsük, hogyan birkóznak meg a szervezetek az AI-modellek kifürkészhetetlenségével, amikor magyarázhatósági igényekkel szembesülnek. A probléma szociotechnikai jellege az esetszervezetnél végzett kutatási projekt korai szakaszában vált nyilvánvalóvá. Megfigyeltük, hogy a szervezet társadalmi oldalát (emberek, folyamatok és szervezeti struktúrák) szinergikusan integrálni kell a szervezet technikai elemeivel (információs technológia és mesterséges intelligencia rendszerek), ha a szervezet ki akarja használni a mesterséges intelligencia modellek szélesebb körét, beleértve a rendelkezésre álló kifürkészhetetlen modellek némelyikét is. Ez a törekvés kétféle célt, a megmagyarázhatósági és a teljesítményorientált célokat foglalja magában, amelyek a mesterséges intelligencia megvalósítása esetében egymásnak ellentmondó követelményeket támasztanak. Itt Sarker et al. (2019) az információs rendszerek megvalósításának instrumentális és humanisztikus eredményeire vonatkozó fogalmaira támaszkodunk, hogy elemezzük a megmagyarázhatóság és a pontosság közötti jól ismert kompromisszumot. A szervezet a nagy teljesítményű mesterséges intelligenciamodellek kifejlesztése során instrumentális irányultságú eredményekre törekedett (jobb teljesítmény és nagyobb hatékonyság), de a humanisztikus eredményekről is gondoskodnia kellett, biztosítva, hogy az ilyen modellek használata ne csökkentse az emberi cselekvőképességet, és ne károsítsa a modellek használata által érintett embereket. Ahogy pontosan feltártuk, hogy a szervezet hogyan kezeli a kívánt eredmények mindkét csoportját, a különböző megközelítések konceptualizálásához a *burkolás (envelopment)* mint megvilágító lencse jelent meg.

Ez a koncepció - a mesterséges intelligencia fejlesztése - a közelmúltban jelent meg, mint potenciálisan hasznos megközelítés a fent leírt magyarázhatósági kihívásokkal való megbirkózásra (Robbins, 2020). Azt sugallja, hogy a képzési adatok gondos ellenőrzésével, a bemeneti és kimeneti adatok megfelelő megválasztásával, valamint egyéb peremfeltételek körültekintő meghatározásával még a kifürkészhetetlen AI számára is lehetővé tehetjük, hogy döntéseket hozzon, mivel ezek a konkrét óvintézkedések kiszámítható burkot emelnek az ágens virtuális manőverezési tere köré. Eddig azonban a burkolást csak néhány kontextusban (pl. autonóm vezetés, Go játék és ruházati cikkek ajánlása) és csak fogalmi szinten mutatták be; így viszonylag korlátozott betekintést nyújtottak a megmagyarázhatósági kihívások kezeléséhez a komplex valós világbeli szervezeteknél. Ennek a hiányosságnak a kiküszöbölése érdekében leírjuk, hogyan gyakorolják az envelopmentet egy úttörő szervezetben, amely megkezdte a mesterséges intelligencia felhasználását a működésében, és megmutatjuk, hogy az envelopment alapvető fontosságú ahhoz, hogy egy szervezet biztonságosan használhasson kiismerhetetlen rendszereket még olyan környezetben is, ahol a megmagyarázhatóságra szükség van. Továbbá elmélyítjük a borítékolás fogalmát azzal, hogy megmutatjuk, hogyan alakul ki a szociotechnikai kölcsönhatásokon keresztül egy komplex szervezeti környezetben. Az itt bemutatott empirikus eredményekkel azt állítjuk, hogy a szociotechnikai burkolás koncepciója széleskörű jelentőséggel bír, és eszközöket kínál számos olyan kihívás enyhítésére, amelyek a fejlett mesterséges intelligencia rendszerek maximális kihasználásának útjában állnak.

2 A szakirodalom áttekintése és az elmélet kidolgozása

Ez a szakasz a szervezeti mesterséges intelligencia megvalósításaiból és azok szociotechnikai alapjaiból már levont tanulságokat tekinti át. Emellett foglalkozunk a jó magyarázatok tulajdonságaival, és részletesebb képet adunk a borítékolási koncepcióról.

2.1 A szervezeti mesterséges intelligencia szociotechnikai megközelítése

Az ML-eszközök új generációinak közelmúltbeli megjelenése és elterjedése újra felébresztette az érdeklődést a szervezeti AI-kutatás iránt (Faraj et al. 2018; Keding 2021; Sousa et al. 2019). Az emberi intelligenciához hasonlóan az AI-t is köztudottan nehéz fogalomként definiálni. Tanulmányunk céljaira követjük Kaplan és Haenlein (2019) definícióját, amikor az AI-t úgy határozzuk meg, mint "a rendszer azon képességét, hogy helyesen értelmezze a külső adatokat, tanuljon ezekből az adatokból, és használja ezeket a tanulságokat meghatározott célok elérésére" (17. o.). A konceptuális munkákat kiegészítve elkezdtek megjelenni a témával kapcsolatos empirikus tanulmányok (pl. Ghasemaghaei, Ebrahimi, & Hassanein, 2018; Salovaara et al., 2019; Schneider & Leyer, 2019). A tanulmányok egyre inkább elmozdították a mesterséges intelligencia kutatásának pozícióját a nagyrészt technikai jellegű kutatásról a *társadalmi* komponenst is magában foglaló perspektíva felé (Ågerfalk, 2020).

Míg a technikai aspektus magában foglalja az információs rendszerek (IS) szemszögét, az IT-infrastruktúrát és a platformokat, addig a társadalmi aspektus az embereket, a munkafolyamatokat, a szervezeti elrendezéseket, valamint a kulturális és társadalmi tényezőket (Sarker et al., 2019). Bár a tudósok megvitatták az olyan kérdéseket, mint az emberek gépekkel való helyettesítése az emberek képességeinek bővítésével szemben (pl. Davenport, 2016; Jarrahi, 2018; Raisch & Krakowski, sajtó alá rendezve), még mindig kevés kritikus empirikus munka vizsgálja az AI szervezeteknél történő telepítésével és irányításával kapcsolatos emberi szempontokat (Keding, 2021).

A mesterséges intelligencia és az automatizált döntéshozatal más formáinak szervezetek általi bevezetésével és használatával kapcsolatos kutatások rámutattak néhány visszatérő mintára. Először is, a mesterséges intelligencia esztelen és ezáltal hibaérzékeny természete szükségessé teszi, hogy a megvalósítás során gondosan ellenőrizzék a mesterséges intelligencia ügynöksége és autonómiája tekintetében. Az ember fontos ellensúlyként szolgálhat ebben az egyenletben (Butler & Gray, 2006; Pääkkönen et al., 2020; Salovaara et al., 2019). Az emberek és a mesterséges intelligencia közötti munkamegosztás és tudás különböző módon rendezhető, így a szervezetek egyensúlyt teremthetnek a merevség és a kiszámíthatóság, valamint a rugalmasság és a kreatív problémamegoldás között (Asatiani et al., 2019; Lyytinen et al., in press). Másodszor, a szervezetek AI-ügynökei számos típusú emberi érdekelt féllel lépnek kölcsönhatásba, amelyek mindegyike sajátos függőséggel rendelkezik a mesterséges intelligenciától és eltérő képességekkel annak működésének megértéséhez (Gregor & Benbasat, 1999; Preece, 2018; Weller, 2019). A tanulmányok azt mutatják, hogy az AI ritkán tekinthető "plug-and-play" technológiának, és hogy az azt bevezető szervezetnek egyértelmű végrehajtási stratégiára van szüksége, amely figyelembe veszi az érdekeltek széles körét (Keding, 2021). Mivel például a mesterséges intelligencia bevezetésének hatása nagymértékben eltér az egyes érdekelt felek között, az érdekelt feleket a tervezés, a bevezetés és a használat folyamatából elválasztó döntések növelik az etikátlan magatartás és a társadalmi szerződések megszegésének valószínűségét, ami gyakran a rendszerek végső kudarcához vezet (Wright & Schultz, 2018).

A szervezeti mesterséges intelligenciával foglalkozó szakirodalom összességében azt mutatja, hogy a szervezetek számára mennyire fontos, hogy egyensúlyt teremtsenek a mesterséges intelligenciával kapcsolatos kockázatok és az elérhető hatékonyságnövekedés között. Ezek a megfontolások azt is mutatják, hogy a szervezeti mesterséges intelligencia alkalmazása jelentős mértékű koordinációt és kölcsönös alkalmazkodást von maga után az emberek és a mesterséges intelligencia között, és így elkerülhetetlenül a szociotechnikai szervezettervezés kérdése (Pääkkönen et al., 2020). A szociotechnikai megközelítés hívei azt állítják, hogy figyelmet kell fordítani mind a technikai artefaktumokra, mind pedig az artefaktumokat társadalmi (pl. pszichológiai, kulturális és gazdasági) kontextusban kifejlesztő és használó egyénekre/kollektívákra (Bostrom et al., 2009; Briggs et al., 2010). Ebből következően a szociotechnikai álláspont elfoglalása instrumentális célok (pl. a modell vagy más modell hatékonysága és pontossága) elérésére irányul.

műtárgy kifejlesztése) és a humanista célok (pl. a felhasználók bevonása és a munkavállalók készségeinek megtartása) egyaránt (Mumford, 2006).

Sarker et al. (2019) áttekintette, hogy a társadalmi és a technikai szempontok milyen bonyolult módon fonódhatnak össze úgy, hogy sem a társadalmi, sem a technikai szempontok nem válnak dominánssá. Megmutatják, hogy ez a kapcsolat igen változatos, és ezt a kölcsönös, valamint a mérséklő hatás, a szociálisnak a technikaiba való beíródása, az összefonódás és az összefonódás példáival mutatják be. A kölcsönös befolyás szempontjából például a technológia és a szervezeti elrendezések úgy tekinthetők, hogy az IS bevezetése során együtt fejlődnek, mivel kölcsönösen kisajátítják egymást (Benbya & McKelvey, 2006). Az imbrikáció szociomateriális perspektívájából viszont az embereket és a technológiákat olyan ügynökségeknek tekintik, amelyek képességei egymásba kapcsolódva rutinokat és más stabilan kialakuló folyamatokat hoznak létre.

2.2 A kifürkészhetetlen mesterséges intelligencia kihívásai

Ahogy a bevezetőben említettük, az összetett mesterséges intelligencia modellek gyakran jobb teljesítményt ígérnek, mint az egyszerűek, de az ilyen modellek általában nem átláthatóak, és kimeneteiket nehéz vagy akár lehetetlen megmagyarázni. A mesterséges intelligencia megmagyarázhatóságáról szóló írások gyakran használják az átláthatóság, az értelmezhetőség és a megmagyarázhatóság egymással összefüggő fogalmait, hogy megpróbálják szétválasztani a probléma szálait. *Az átláthatóság* a mesterséges intelligencia belső műveleteinek nyomon követhetőségére utal - például annak nyomon követésére, hogy milyen utakon keresztül jut el a mesterséges intelligencia a következtetéseihez (Rosenfeld & Richardson, 2019; Sørmo et al., 2005). Ennek ellentéte az átláthatatlanság, a "fekete dobozos" rendszerek tulajdonsága, amelyek elrejtik a döntési folyamatot a felhasználók, sőt néha még a rendszer fejlesztői elől is (Lipton, 2018). A másik két fogalom - az *értelmezhetőség* és a *megmagyarázhatóság* - *a* mesterséges intelligencia kimeneteinek az ember számára való érthetőségére *utal* (pl. Doshi- Velez & Kim, 2017; Miller 2019). A fogalmakat esetenként felcserélve használják (pl. Došilović et al., 2018; Liu et al., 2020), míg néha a szerzők külön definíciókat alkalmaznak. Gyakran az értelmezhetőségnek erős technikai konnotációi vannak, míg a megmagyarázhatóság emberközpontúbb természetű, ezért inkább szociotechnikai irányultságú fogalom.

A hagyományosabb mesterséges intelligencia modellek közül sok, például a lineáris regresszió, amely csak korlátozott számú ismert bemeneti változót kezel, és a döntési fák, amelyek képesek megjeleníteni a követett ha-akkor szekvenciát, magyarázhatónak tekinthetők. A mai mesterséges intelligencia modellek közül azonban egyre több olyan összetett, hogy a megmagyarázhatóság gyakorlatilag lehetetlenné válik. Ha például egy hagyományos döntési fa modellt a gradiens növelésnek nevezett gépi tanulási technikával "felturbózunk", a teljesítménye javul, de a viselkedése sokkal nehezebben magyarázhatóvá válik. Más példák a nagy pontosságú modellekre, amelyek nem magyarázhatók meg, a mély- és a

rekurrens neurális hálózatok, összetetten rétegzett számítástechnikai rendszerek, amelyek felépítése hasonlít az agyi neuronok biológiai hálózataihoz. Ezután az ember *kifürkészhetetlennek* tartja őket (Dourish, 2016; Martin, 2019), utalva azokra a helyzetekre, amikor a rendszer összetettsége meghaladja az átfogó elemzés gyakorlati eszközeit. A Google-nál nemrégiben kifejlesztett, nyílt tartományban működő chatbot, amelynek mély neurális hálózatában 2,6 milliárd szabad paraméter van (Adiwardana et al., 2020), szélsőséges példája egy olyan mesterséges intelligencia rendszernek, amelynek belső működése az ember számára még akkor is kifürkészhetetlen, ha átlátható.

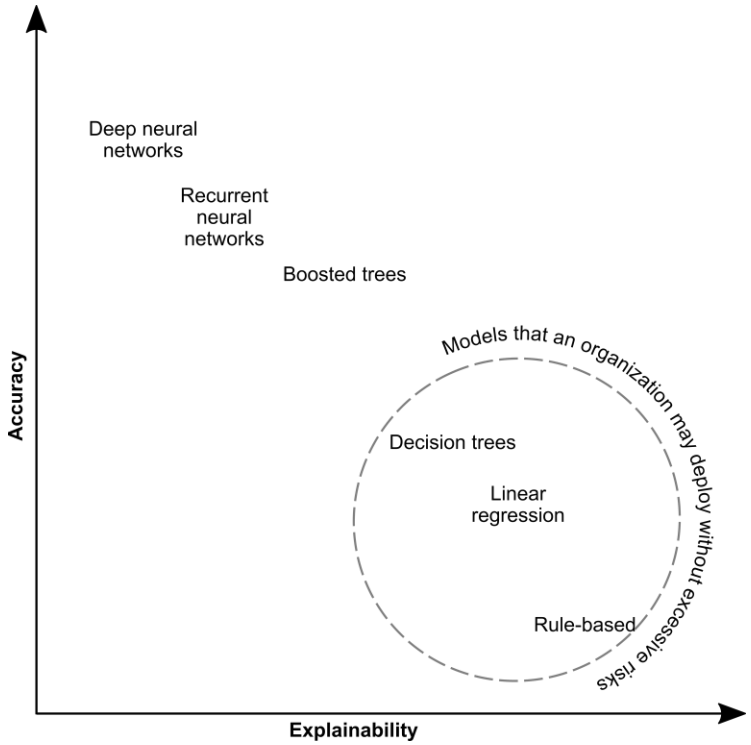
A kifürkészhetetlen rendszerek korlátlan használata problémás lehet. Az ilyen rendszerekkel interakcióba lépő emberek nem tudják validálni, hogy a rendszer által hozott döntések megfelelnek-e a valós világ követelményeinek és betartják-e a jogi vagy etikai normákat (Rosenfeld & Richardson, 2019). A kérdés messze nem akadémikus, elvégre a kiismerhetetlen rendszerekre való támaszkodás a döntéshozatalban szisztematikus torzításokhoz vezethet, amelyek teljesen láthatatlanok a rendszerrel interakcióban lévő vagy a rendszer által érintett emberek számára (Došilović et al., 2018).

Ennek következtében a mesterséges intelligencia-rendszereket telepíteni szándékozó szervezetek a *magyarázhatóság és a pontosság közötti kompromisszummal* szembesülnek (Došilović et al., 2018; Linden et al., 2019; London, 2019; Martens et al., 2011; Rosenfeld & Richardson, 2019). Egyrészt a nagyobb rugalmassággal rendelkező komplex modellek, például a mély neurális hálózatok gyakran pontosabb előrejelzéseket adnak, mint az egyszerű modellek, például a lineáris regresszió vagy a döntési fák. Másrészt az egyszerű modellek általában könnyebben értelmezhetők és magyarázhatók az emberek számára. A megmagyarázhatóság és a pontosság között fennállónak tűnő kompromisszum arra kényszeríti a tervezést, hogy az egyiket a másikkal szemben előnyben részesítse: egy olyan szervezetnek, amely csökkenteni kívánja a megfejthetetlen mesterséges intelligenciával kapcsolatos kockázatokat, magas fokú megmagyarázhatósággal rendelkező mesterséges intelligencia modellekkel kell beérnie. Az 1. ábra ezt a kompromisszumot szemlélteti, Linden et al. (2019) és Rosenfeld és Richardson (2019) ábrázolásait követve.

A közelmúltban bevezetett egyik megközelítés a fekete dobozos rendszerekből eredő kockázatok kezelésére a burkolás. A magyarázhatóság és a pontosság közötti kompromisszum kezelésében rejlő lehetőségek elismeréseként a következő szakasz a kutatók által ezzel a megközelítéssel kapcsolatban tett javaslatokat vizsgálja meg.

2.3 Envelopment

Amint fentebb említettük, a borítékolást (Floridi, 2011; Robbins, 2020) a szervezeti AI-fejlesztés területének vizsgálatakor megfelelő érzékelési fogalomként azonosítottuk. Eredeti kontextusában a robotikában a *munkaburok* "a robotkéz vagy munkaeszköz maximális kiterjedését vagy hatótávolságát minden irányban reprezentáló pontok halmaza" (RIA Robotics Glossary, 73; idézi Scheel, 1993, 30. o.).



1. ábra. A megmagyarázhatóság és a pontosság közötti kompromisszum

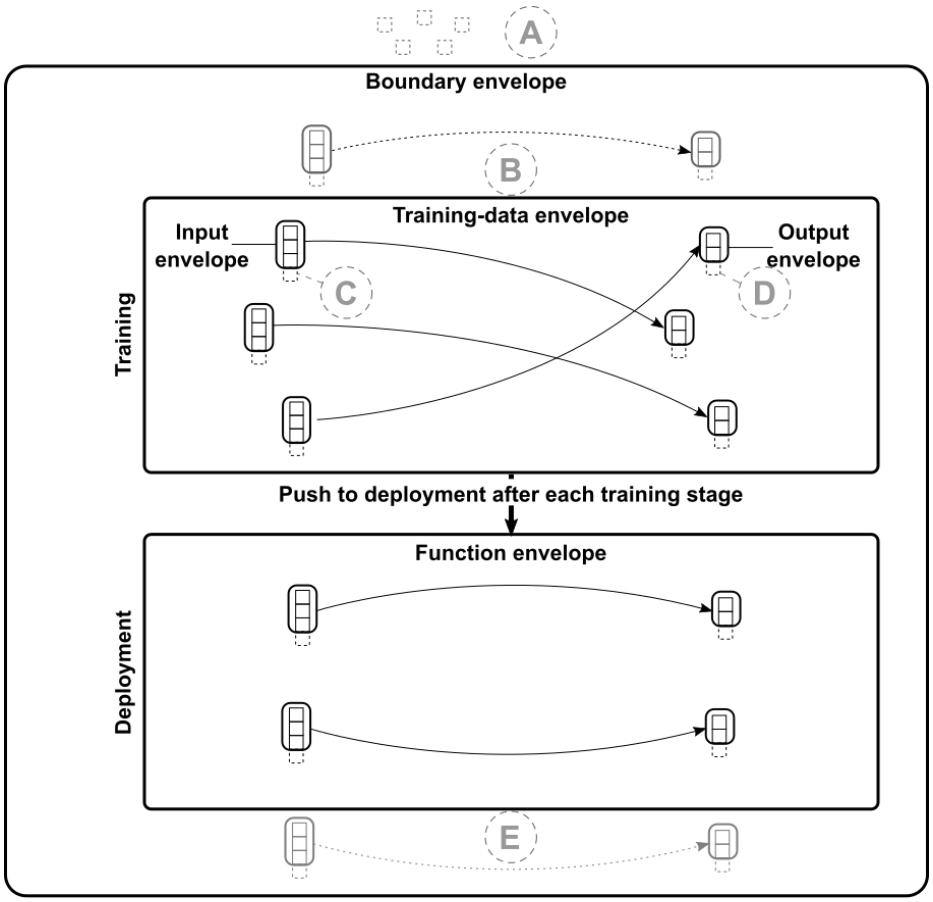
A robotok munkaterületei, amelyeket gyakran árnyékolt régióként ábrázolnak a gyárak alapterületi térképein és csíkos területként a gyárak padlóján, gyakorlati megoldást jelentenek az úgynevezett "szükséges változatosság elvének" (Ashby, 1958) teljesítésére - azaz annak a követelménynek a teljesítésére, hogy a robot logikájának állapotainak száma nagyobb legyen, mint a környezeti állapotok száma, amelyekben működik. Ha egy robot olyan környezetben cselekszik, amelynek komplexitása meghaladja a felfogóképességét, akkor veszélyt jelent a környezetére. A munkakörülmények - olyan területek, ahová más szereplők nem lépnek be - garantálhatják, hogy a robot fizikai környezete kellően leegyszerűsödjön (azaz a környezet lehetséges állapotainak száma kellően lecsökkenjen). Ezzel a módosítással a robot képes kezelni azokat az állapotokat, amelyeket még irányítani kell, és ezáltal teljesül a szükséges változatosság elve. A fizikai paraméterek mellett a robot burkolt változatosságát időbeli küszöbértékek, szükséges képességek/felelősségek és elfogadott feladatok segítségével is meg lehet határozni (McBride & Hoffman, 2016, 79. o.). Ezek a paraméterek dinamikusak: amikor a robot új problémákkal szembesül, a burkoló paraméterek kiigazításra kerülnek, hogy alkalmazkodjanak ahhoz, amit a szükséges változatosság most magával hoz (81. o.).

Kutatásunk annak a munkának a folytatása, amelyben ezt a koncepciót olyan esetekre alkalmaztuk, amelyekben emberek és mesterséges intelligencia-ügynökök által végzett nem fizikai munka szerepel. Ebben az összefüggésben a burkolat nem fizikailag meghatározott, hanem az információfeldolgozás területére vonatkozik. Ettől a területváltozástól függetlenül továbbra is szükség van az emberi partnerrel való együttműködésre, aki fenntartja a

a borítékot, és így garantálja a mesterséges intelligencia működésének biztonságát és helyességét (Floridi, 2011). Emellett továbbra is fennmarad a szükséges változatosság alapelve, ami azt jelenti, hogy a mesterséges intelligenciát nem szabad olyan feladatokra használni, amelyeket nem tud elsajátítani, és nem szabad a feladatok szempontjából irreleváns adatokkal betanítani. Az ilyen nemkívánatos hatások - az 1. ábrán a "túlzott kockázatok" - számos formában jelentkezhetnek, többek között hibás bemenet-cselekvés leképezésekben, olyan etikai dilemmákban, amelyeket egy mesterséges intelligencia-ügynöknek nem szabadna magától megoldania, valamint elfogultságot mutató viselkedésekben (pl. Robbins, (2020). Még ha az ilyen kockázatok megvalósulása nem is rontja a pénzügyi végeredményt vagy a műveletek hatékonyságát, problematikus humánus eredményeket eredményezhet. Például egy mesterséges intelligenciával működő rendszer, amely az álláspályázatok feldolgozásával azonosítja a legígéretesebb jelölteket, növelheti a HR-osztály hatékonyságát, és következetesen azonosítja a pozíció követelményeinek megfelelő jelölteket. Ugyanakkor a rendszer következetesen diszkriminálhatja a jelentkezők bizonyos csoportjait, akik egyébként megfelelnének a követelményeknek, mert a mögöttes modellben előítélet van. Ilyen forgatókönyvek esetén a mesterséges intelligenciával kapcsolatos intézkedések - legalábbis rövid távon - nem befolyásolják a vállalat eredményét, de ettől függetlenül problémásak lehetnek.

A borítékolás többféle módszerrel fejleszthető. A 2. ábra a Robbins (2020) által megfogalmazott öt módszer értelmezését mutatja be. Az alábbiakban összefoglaljuk őket, majd a tanulmányunkkal kapcsolatban építünk rájuk. A *Boundary Envelopes a* legáltalánosabbat képviseli a

borítékolási módszerek. A burkológörbe kijelöli, *hogy* a mesterséges intelligencia *hol* működik - például csak a jó fényviszonyok között lefényképezett emberi arcokról készült képeket elemzi. Egy ilyen módon burkolt mesterséges intelligenciamodell nem találkozik más feladatokkal, mint a számára gondosan kijelöltekkel (A feltétel a 2. ábrán). Robbins (2020) egy robotporszívó tervezését veszi példának. Határfelületi burkolási mechanizmusa azt jelenti, hogy a robot nem kell, hogy képes legyen elkerülni olyan veszélyeket, amelyek a beltéri lakótérben soha nem léteznek (pl. víztócsák). A határok behatárolásának előnye, hogy a mesterséges intelligenciának nem kell olyan módszereket beépítenie, amelyekkel felismerhető, hogy az ágens olyan forgatókönyvekben működik-e, amelyek túlmutatnak azon a képességén, hogy felfogja a környezetet (azaz a szükséges változatosságot).



Legend:

An input or output vector of data. One vector element (dashed gray line) has been enveloped out and is not used in the model. Rectangles with bold stokes denote envelopes.

Examples of what the envelope scope excludes:

A Events and states-of-affairs in the world that the model does not need to "know" about. [Boundary]

B Input–output pairs that could be used in training data but are suspected of bias, errors, or represent cases for which not enough data exists yet and the model should not be allowed to learn from. [Training-data envelope]

C Input sources that would provide low-quality information. [Input envelope]

D Outputs that a model could provide but that are biased, not needed, or redundant. [Output envelope]

E Purposes for which the trained model will not be used (e.g., for ethics reasons), even if it would be capable of accurate performance. [Function envelope]

2. ábra. A Robbins (2020) által javasolt AI Envelopment módszerek szemléltetése

A többi borítékolási módszer közül három olyan van, amely arra a fogalomra vonatkozik*, hogy milyen tartalmat* manipulál az AI (Robbins, 2020). Az első ezek közül a *képzési adatok burkolása, amely a* helyes bemeneti-kimeneti leképezések kurátori kezeléséhez kapcsolódik, amelyekkel az AI-modell betanításra kerül. Robbins szerint az előítéletek és más reprezentativitási problémák ("B" a 2. ábrán) különösen valószínűsítik a társadalmi sztereotípiák terjedését vagy fenntartását, ha a borítékot nem kezelik megfelelően. A *bemeneti borítékok viszont a* mesterséges intelligencia bemeneteinek technikai részleteivel foglalkoznak. Robbins példájában például egy ajánló AI különböző időjárási és felhasználói adatokat (pl. hőmérséklet, valós idejű időjárási állapot és a felhasználó naptára) használ fel ruházati ajánlások (pl. esőkabát viselésének javaslata) készítéséhez. A jó eredményekhez az adatoknak jó minőségű, zajmentes és megfelelő szemcseméretű forrásokból kell érkezniük. A bemeneti burkolás a bemeneti csatornákat azokra korlátozza, amelyek e tekintetben megfelelnek a megfelelő kritériumoknak, és megakadályozza, hogy a rosszul értelmezett források befolyásolják a modell viselkedését. A harmadik burkoló módszer a "mi" kategóriában a *kimeneti burkolatok használata*. Ezek határozzák meg a mesterséges intelligencia működésének hatókörén belül végrehajtható műveletek körét. Egy autonóm módon közlekedő autó esetében a kimenetek a következők lehetnek: gyorsítás, a kerekek elforgatása és fékezés. Még ha a gyorshajtás technikailag lehetséges és néha hasznos is lenne, kockázatot jelent az utasokra és a többi közlekedőre nézve. Ezért ez a kimenet ki van zárva az autonóm autó cselekvéseiből. A 2. ábrán a "C" és a "D" ábra a fent leírt bemeneti és kimeneti fejlesztési módszereket szemlélteti.

Az ötödik és egyben utolsó módszer, a *funkcióburok* használata arra a kérdésre keresi a választ, hogy *miért* létezik a mesterséges intelligencia, és milyen célok és etika előmozdítására tervezték. A burkolásnak ezt a kategóriáját arra alkalmazzák, hogy korlátozzák a mesterséges intelligencia rosszindulatú vagy egyéb problémás célokra való használatát, még azokban az esetekben is, amikor az helyesen működik. Például az olyan társalgási otthoni asszisztensek, mint az Echo vagy az Alexa funkcióit a magánélet megsértésének elkerülése érdekében csak a háztartási tevékenységek egy szűk körére korlátozzák (Robbins, 2020). A funkciók ilyen kiszűrése a 2. ábrán "E"-vel van jelölve.

Robbins azt javasolja, hogy a rendelkezésre álló borítékolási módszerek ilyen sokféleségével vagy leküzdhetünk néhány, a fekete dobozos mesterséges intelligenciával kapcsolatos problémát, vagy semlegesíthetjük azok hatásait. Munkánkat tehát a borítékolás koncepciója alapozza meg, és annak alkalmazhatóságát komplex és kialakulóban lévő szociotechnikai környezetben vizsgáljuk. Különösen azt állítjuk, hogy az emberek fontos szerepet játszanak egy mesterséges intelligencia-ügynök burkolásában és annak megszervezésében, mivel arra törekszünk, hogy az AI ne szembesüljön olyan feladatokkal, amelyeket nem tud feldolgozni vagy helyesen értelmezni - ha a problémák meghaladják a szükséges változatosságot (pl. Salovaara et al., 2019). Ezután beszámolunk az esettanulmányunkról.

3 Az esettanulmány: Gépi tanulás kormányzati környezetben

Annak vizsgálatára, hogy egy szervezet hogyan kezelheti a megmagyarázhatósági kihívásokat, egy feltáró esettanulmányt végeztünk egy kormányzati ügynökségnél, amely több ML-projekten keresztül aktívan törekszik a mesterséges intelligencia alkalmazására. Olyan szervezetet választottunk, amely kiterjedt képességekkel rendelkezik az AI/ML-eszközök fejlesztésére, valamint elkötelezett az elszámoltathatóság és a megmagyarázhatóság iránt.

3.1 A vizsgálat helyszíne

A Dán Gazdasági Hatóság (DBA) a dán Ipari, Üzleti és Pénzügyi Minisztérium alá tartozó kormányzati szerv. Körülbelül 700 alkalmazottat foglalkoztat, székhelye Koppenhágában van, és Silkeborgban és Nykøbing Falsterben is működik egy-egy részlege. A hatóság feladata az üzleti élethez kapcsolódó alapvető feladatok széles köre, amelyek a vállalkozások növekedési potenciáljának fokozása köré csoportosulnak Dánia-szerte. A DBA tartja fenn a VIRK digitális platformot, amelyen keresztül a dán vállalatok üzleti dokumentumokat nyújthatnak be, és amely lehetővé teszi a DBA számára az online cégnyilvántartás vezetését (amely körülbelül 809 000 vállalatot tartalmaz, összesen körülbelül 812 000 regisztrációval, amelyek együttesen évente körülbelül 292 000 éves nyilatkozatot nyújtanak be). A DBA karbantartási és végrehajtási feladatai olyan törvényekkel kapcsolatosak, mint a dán társasági törvény, a pénzügyi kimutatásokról szóló törvény, a könyvelési törvény és a kereskedelmi alapítványokról szóló törvény. A múltban a DBA együttműködött az Early Warning Europe (EWE) hálózattal is - amely a vállalatok és vállalkozók megsegítésére jött létre Európa-szerte -, hogy támogatási mechanizmusokat dolgozzon ki a bajba jutott vállalatok számára. A tanulmányunkban elemzett ML-projektek a DBA alapvető feladataihoz kapcsolódnak - például a VIRK-felhasználók viselkedésének megértéséhez, valamint a cégbejegyzések és éves beszámolók hibák és csalásra utaló jelek szempontjából történő ellenőrzéséhez.

Az ML használatának ötlete a DBA-nál 2016-ban született. Az ügynökség mesterséges intelligenciával kapcsolatos piackutatásba kezdett, amely 2017-ben több adattudományi projektben és a Machine Learning Lab (innentől kezdve "ML Lab") létrehozásában csúcsosodott ki. Az ML Lab létrehozásának egyik mozgatórugóját az adta, hogy a DBA által feldolgozott különböző típusú dokumentumok mennyiségének óriási növekedése volt. Ahelyett, hogy külső tanácsadókat vett volna igénybe és támaszkodott volna rájuk, a DBA úgy döntött, hogy saját adatmérnököket és adattudósokat vesz fel. Ennek a házon belüli megközelítésnek a fő okai a költséggazdálkodási megfontolások és az a törekvés voltak, hogy a releváns tudást az ügynökségen belül tartsák. Az ML-megoldások belső létrehozása olyan technológiák kombinálásával, mint a Neo4j gráfadatbázis-kezelés, a Docker konténerek és a Python, jobban illeszkedik a szervezethez, mint a kereskedelmi forgalomban kapható kész megoldások. Emellett az ML Lab szerepe

nagyrészt a koncepciót igazoló modellekkel kapcsolatos kísérletezésre és fejlesztésre korlátozódik. Ha egy megoldás hasznosnak bizonyul és megfelel a meghatározott minőségi kritériumoknak, akkor a bevezetését külső tanácsadó cégekre bízzák, amelyek aztán a modellt a gyártásban alkalmazzák. Ez a döntés elsősorban a DBA-kultúrán alapult, amelyben a szállítók vállalják a felelősséget a kódjukkal kapcsolatos támogatási és karbantartási funkciókért: az ML-modellek ugyanolyan irányítást követnek, mint a DBA-n belül más IT-projektek.

Ezért a DBA ML-hez kapcsolódó műveletei két fő egység között oszlanak meg: egy fejlesztési egység (az ML Lab) és egy végrehajtási egység (külső tanácsadók) között. Az ML Lab szerepe az, hogy szorosan együttműködjön a szakterületi szakértőkkel (a továbbiakban "esetmunkások") a funkcionális prototípusok kifejlesztése érdekében, a koncepció bizonyításának részeként. A laboratórium fő célja annak bizonyítása, hogy az esettanulók által azonosított problémák megoldhatók az ML segítségével. A koncepció bizonyítása és a dokumentáció, például az értékelési terv együttesen képezi az alapját annak, hogy a DBA irányítóbizottság döntést hozzon arról, hogy a modellt továbbítja-e a megvalósítási egységnek. A folyamat különböző részeiért különböző érdekelt felek felelősek. Az ML Lab felelős a prototípus kifejlesztéséért, és az eseti dolgozók a prototípus kifejlesztése során a labor munkatársai számára a szakterületi ismereteket biztosítják. Az esetmunkások az ML-modellek működési helyességéért is felelnek, mivel az egyes modellek értékelésével és szükség szerinti átképzésével is megbízzák őket. Az irányítóbizottság ezután dönt arról, hogy mely modellek és mikor kerüljenek be a sorozatgyártásba. Végül az implementációs egység felelős a modell megvalósításáért és a technikai karbantartás felügyeletéért.

3.2 Adatgyűjtés

A DBA-nál végzett interjúk és megfigyelések szolgáltak fő adatforrásként. Célzott mintavételt alkalmaztunk (Bernard, 2017), és az alábbi kritériumok alapján választottuk ki az esetszervezetet. A szervezetnek fejlett AI- és ML-képességekkel kellett rendelkeznie, mind az erőforrások, mind a know-how tekintetében. Elkötelezettnek kellett lennie továbbá a megmagyarázható rendszerek fejlesztése iránt. Végül a kutatóknak hozzáférésre volt szükségük az AI/ML projektekhez, a kapcsolódó folyamatokhoz és a releváns érdekeltekhez. Az utolsó kritérium különösen fontos volt ahhoz, hogy szélesebb perspektívát kapjunk a projektekről, és hogy ellenőrizni tudjuk az informátorok által megfogalmazott magyarázhatósági állításokat. A DBA mindezen kritériumoknak megfelelt.

A DBA-hoz való hozzáféréshez az ismert szponzor megközelítést használtuk (Patton, 2001): a DBA-nál a szervezeten belüli ML-kezdeményezésekkel foglalkozó felsővezetőhöz volt hozzáférésünk, aki segített nekünk az interjúk megszervezésében az adatgyűjtés korai szakaszában. A vezető legitimitására és hitelességére támaszkodva már a kezdetektől fogva megalapoztuk legitimitásunkat és hitelességünket a DBA-n belül (Patton, 2001). Ezenkívül az egyik szerzőnek munkakapcsolata volt a szervezetnél a

műveleti szinten, lehetővé téve számunkra, hogy az adatgyűjtési munka további szakaszában interjúkat szervezzünk. Ez segített abban, hogy kölcsönös bizalmat alakítsunk ki az informátorokkal, és megakadályozta, hogy a felső vezetés ügynökeinek tekintsenek bennünket.

Az adatokat egy négylépcsős iteratív folyamat keretében gyűjtöttük és elemeztük (lásd az 1. táblázatot), amelyben a fázisok átfedték egymást, és a korábbi szakaszok tájékoztatták a későbbi szakaszokat. Az elit torzítás elkerülése érdekében arra törekedtünk, hogy a DBA alkalmazottak széles körével, a hierarchia több szintjén, különböző szolgálati idővel rendelkező munkatársakkal készítsünk interjút (Miles et al., 2014; Myers & Newman, 2007). Az 1. szakasz feltáró jellegű volt. Célja a kutatási együttműködés kialakítása és a DBA jelenlegi és jövőbeli ML-projektjeiről és elképzeléseiről alkotott kép kialakítása volt adattudományi és esettudományi szempontból. A második fázis célja a DBA különböző ML-projektjeinek és az érintett szereplőknek a mélyebb megértése volt. Ebben a fázisban az ML Laborra és annak a projektekben betöltött szerepére és felelősségi körére, valamint az ML-hez kapcsolódó magyarázhatóságra összpontosítottunk. Ezután, a 3. fázisban interjút készítettünk az ML Lab összes alkalmazottjával, valamint két, a laborral szoros együttműködésben tevékenykedő eseti munkatárssal. Az utolsó fázis az elemzésünkből származó értelmezések validálását és a labort támogató technikai infrastruktúrába való további betekintést jelentette.

Félig strukturált interjúkat készítettünk minden fázisban, amelyekre 2018 augusztusától 2020 októberéig került sor. A kezdeti benyomások fontosak a kutatók és az informátorok közötti bizalom kialakításához (Myers & Newman, 2007); ezért mindig úgy mutatkoztunk be, mint egy tudományos tanulmányt végző pártatlan kutatócsoport. Minden egyes interjú elején elmagyaráztuk a tanulmány általános célját és az adott informátor(ok) részvételre való kiválasztásának okait. Minden informátornak anonimitást és titoktartást ígértünk, és kifejezett beleegyezést kértünk az interjúk rögzítéséhez. Elmagyaráztuk továbbá, hogy a beleegyezést bármikor visszavonhatják az interjú során vagy azt követően, egészen a kutatási cikk végleges közzétételéig. Gondoskodtunk arról, hogy az informátorok minden, az eljárással kapcsolatban felmerült aggályára kitérjünk, és minden kérdésre válaszoltunk.

Az interjúk angol nyelven zajlottak, az egyik szerző, aki dán anyanyelvű, mindegyiken jelen volt, és szükség esetén tisztázta a terminológiát. Emellett az informátoroknak lehetőségük volt arra, hogy dánul beszéljenek, ha úgy kívánták. Az angol mint elsődleges nyelv kiválasztása annak figyelembevételével történt, hogy a kutatócsoport legtöbb tagja nem beszélt dánul, míg az összes informátor nagyfokú angol nyelvtudással rendelkezett. Bár felismertük az interjúk olyan nyelven történő lefolytatásának lehetséges hátrányait, amely nem az interjúalanyok anyanyelve, elfogadtuk a fennmaradó kockázatot annak érdekében, hogy a teljes kutatócsoport részt vehessen az adatgyűjtési folyamatban és az adatelemzésben. Minden interjút hangfelvételen rögzítettünk és átírtunk, így 167 006 szónyi szöveget kaptunk.

1. táblázat. Az adatgyűjtés négy fázisa

Fázisszám, téma és dátumtartomány	**Módszer és időtartam**	**Az informátor álneve és szerepe**	**Az eredmények középpontjában**
1. ML projektek összességében, 2018. augusztus-szeptember	Csoportos interjú (105 perc)	James (ML Lab csoportvezető / vezető adattudós); Mary (vezető tanácsadó)	A DBA feladatai; szervezet szerkezet
2. ML Lab funkciók, 2018. október - január 2019	Személyes interjú (90 perc)	James	A magyarázhatóság szerepe az ML-projektekben; a feladatok elosztása az érdekelt felek között (az ML-laboratórium, a végrehajtási egység és az esetmunkások).
	Csoportos interjú (83 perc)	David; John (mindketten a Korai Figyelmeztető Rendszer külső esetfelelősei)	
	Személyes interjú (70 perc)	Daniel (belső ügyintéző)	
	Személyes interjú (59 perc)	Steven (az ML Lab adattudósa)	
	Személyes interjú (51 perc)	Mary	
	Személyes interjú (116 perc)	James	
3. Magyarázhatóság az ML-projektekben, 2019. szeptember	Személyes interjú (51 perc)	Steven	A megmagyarázhatósági kérdések gyakorlati eszközei; a modellfejlesztés szociotechnikai környezete
	Személyes interjú (54 perc)	Thomas (az ML Lab adattudósa)	
	Személyes interjú (50 perc)	Linda (az ML Lab adattudósa)	
	Személyes interjú (48 perc)	Michael (az ML Lab adattudósa)	
	Személyes interjú (52 perc)	Mark (az ML Lab adattudósa)	
	Személyes interjú (53 perc)	Joseph (az ML Lab adattudósa)	
	Személyes interjú (54 perc)	Jason (az ML Lab egyik csoportvezetője)	
	Személyes interjú (48 perc)	Susan (az ML Lab adattudósa)	
	Személyes interjú (62 perc)	William (belső ügyintéző)	
	Személyes interjú (54 perc)	Daniel	
4. Az elemzésből származó értelmezések ellenőrzése, 2019 decembere és 2020 októbere között.	Személyes interjú (55 perc)	Jason	Az értelmezések validálása az interjúk visszajelzései és egy projekt-sablonok segítségével történő feltérképezést magában foglaló értékelési gyakorlat révén.
	Értékelési gyakorlat (idő N/A)	Steven; Mary; Thomas; Linda; Michael; Mark; Joseph; Jason; Susan	
	Személyes interjú (27 perc)	Jason	
	Személyes interjú (32 perc)	Steven	
	Személyes interjú (49 perc)	Daniel	

Az interjúk mellett résztvevő megfigyelést és dokumentumelemzést is alkalmaztunk. A dánul beszélő szerző által vezetett kézzel írt terepnaplók háttérinformációkat szolgáltattak. Ezek 2017 szeptemberéig nyúlnak vissza, amikor a DBA-nál kapcsolatba került az ML-lel. A külső tanácsadóként, majd a Koppenhágai Informatikai Egyetem és a DBA által egyenlő arányban finanszírozott közös PhD-hallgatói munkát lefedő naplóanyag megfigyelésekből, feladatleírásokból és találkozókon készített feljegyzésekből áll. A naplók a kutatási időszak teljes időtartamára kiterjedtek, beleértve azt az időszakot is, amikor a legtöbb ML-projekt vagy a fejlesztés nagyon korai szakaszában volt, vagy még el sem kezdődött. A DBA-nál körülbelül minden második munkanapot figyelembe véve, a

a doktorandusz megfigyelései reális képet adnak az esetet vizsgáló szervezet mindennapi munkájáról. A terepnaplókat memóriatámogatásként, az interjúadatok hiányosságainak pótlására, valamint a kulcsinformátorokra, a szervezeti struktúrára, a szervezeti folyamatokra és a munkamódszerekre vonatkozó alapvető információk referenciájaként használtuk. Ezenkívül a naplók segítettek megerősíteni az informátorok egyes állításait. Hasonlóképpen, a dokumentumelemzés az érdeklődés teljes időtartamára kiterjedt. Ez a munka magában foglalta a DBA Jira rendszeréből, egy projektmenedzsment eszközből kinyert dokumentáció és felhasználói történetek elemzését. A dokumentumelemzés kiterjedt a DBA Git-tárának (a verziókezelő rendszerben használt) elérésére és annak ellenőrzésére is, hogy melyik modell volt

minden egyes projektben. Ezenkívül az együttműködő doktori kutatónak hozzáférése volt egy személyes e-mail fiókhoz a szervezetnél, és kereshetett a régi beszélgetésekben, illetve újakat indíthatott, ha az ML-projektek során hozott döntések további magyarázatot igényeltek. Végül a szerzők elemzése során felmerülő értelmezések ellenőrzése érdekében megkértük az ML Lab adattudósait, hogy egy értékelési gyakorlat keretében a szerzőkkel együtt töltsék ki az egyes ML-projektek vázlatos dokumentumát. Ez a gyakorlat egy *input-ML-modell-kimenet* keretrendszert hozott létre, amely lehetővé tette számunkra az ML-projektek alapjainak ellenőrzését és egységes projektleírások létrehozását, amelyek jellemzik például a modellbe táplált adatokat, az alkalmazott ML-modell típusát és az előállított kimenet jellegét. Az A. függelék összefoglalja ezt a keretrendszert.

3.3 Adatelemzés

Összességében elemzési megközelítésünk abduktívnak tekinthető: induktívnak indult, de később egy elméleti lencse tájékoztatta, amely megfelelő érzékenyítő eszközként jelent meg (Sarker et al., 2018; Tavory & Timmermans, 2014). Az összes interjúadatot három szakaszban kódoltuk, a megalapozott elmélet kevésbé eljárásorientált változataiból átvett kódolási és elemzési technikákat alkalmazva (Belgrave & Seide, 2019; Charmaz, 2006). A gyakorlatban ez azzal járt, hogy a kezdeti fogalmak azonosításához állandó összehasonlító elemzésre támaszkodtunk. Az adatgyűjtési és elemzési folyamatok kölcsönösen integrálták egymást (Charmaz, 2006), folyamatosan átvezetve minket a konkrét interjú és az esetszervezet tágabb kontextusa között (Klein & Myers, 1999). Később a felmerülő fogalmakat magasabb szintű kategóriákhoz kapcsoltuk. A megalapozott elmélet elemeinek kvalitatív adatelemzésre való felhasználására irányuló megközelítésünk és a korábbi IS-tanulmányokban (pl. Asatiani & Penttinen, 2019; Sarker & Sarker, 2009) kialakított módszerek között hasonlóságok figyelhetők meg.

A kódolás három szakasza fogalmakat (elsőrendű konstrukciók), témákat (másodrendű konstrukciók) és összesített dimenziókat (lásd a C függeléket) eredményezett, párhuzamosan a Gioia, Corley és Hamilton (2013) által javasolt struktúrával. Az első szakaszban nyílt kódolást végeztünk, amelynek kódjai teljes mértékben az adatainkban alapultak. Ez bekezdésről bekezdésre történő kódolást jelentett, közvetlenül az informátorok diskurzusából vett in vivo kódokat használva (Charmaz, 2006), a kódolók minimális értelmezésével. Például a következő részlet: "Lenne egy irányítási küszöb. Valójában nem. Ennél a modellnél lenne valami általunk meghatározott iránymutatás, igen. És aztán az esetkezelők szabadon mozgathatják felfelé és lefelé" - két kódot kaptak: "az esetkezelői ellenőrzési küszöbértékek" és "irányítási küszöbérték". *" A* szerzők közül ketten egymástól függetlenül végezték a nyílt kódolást, majd a két kódkészletet újra átnézték, összehasonlították és finomították. A fogalmilag hasonló kódokat összevonták a fogalomkészletbe.

A második szakaszban elemeztük a nyílt kódolás eredményeit, és elkezdtük keresni a felmerülő témákat. A nyílt kódok és az interjúk leiratai között iteráltunk, és az adatokat több fogalmat összekötő tágabb témák (axiális kódolás) érdekében kódoltuk. Bár ezek a témák magasabb szinten voltak, mint az első szakasz in vivo kódjai, még mindig szilárdan megalapozottak voltak az adatokban. Minden szerző részt vett ebben a szakaszban, amely az azonosított kódok összehasonlításában és konszolidálásában csúcsosodott ki a másodrendű konstrukciók - a témák - előállítása érdekében.

A harmadik szakaszban elméleti kódolást alkalmaztunk az adatainkra. E kifejezés ellenére e szakasz célja nem egy konkrét elmélet érvényesítése volt. Inkább a DBA megközelítéseit akartuk rendszerezni a megmagyarázható mesterséges intelligencia kihívások megoldására, ahol az átlátható rendszer kiépítése nem volt opció. Ehhez Robbins (2020) borítékolási kerete szolgált érzékenyítő lencseként, amely segített az elemzés második szakaszában felmerült témák rendszerezésében. A döntés adatvezérelt volt - nem számítottunk arra, hogy ilyen erős fókuszt találunk a borítékolásra az eseti szervezetnél, de az elemzés első két szakasza induktív módon feltárta, hogy a DBA stratégiája inkább hasonlít a borítékolásra, mint egy olyan módszerre, amellyel a DBA megpróbálja garantálni a magyarázhatóságot az összes AI-modell megvalósításában. A munka ezen szakaszában minden szerző részt vett, egymástól függetlenül végezték a kódolást. Ezután a kódokat összeállították, összehasonlították és egyetlen kódkészletté szintetizálták.

4 Találatok

Megállapításaink a DBA ML Lab nyolc mesterséges intelligencia-projektben végzett munkájából származnak, amelyeket itt könyvvizsgálói nyilatkozat, csőd, cégbejegyzés, föld és épületek, személyazonosság ellenőrzése, ajánlás, ágazati kód és aláírás néven jelölünk (a projekt részleteit lásd az A. függelékben). Bár minden projektnek külön célja volt, mindegyiknek az volt a célja, hogy támogassa a DBA-nak mint kormányzati üzleti hatóságnak a társadalomban betöltött szerepét. E dokumentum megírásának időpontjában e projektek közül sok már beindult és folyamatos használatba került. A DBA-ra nagy nyomás nehezedett, hogy rendkívül hatékony legyen, ugyanakkor a nyilvánosság szemében átlátható és megbízható szereplő maradjon, és a mesterséges intelligencián alapuló eszközök hatékony alternatívát jelentettek a rendkívül erőforrás-igényes, teljesen emberi alapú adatfeldolgozással szemben. Ugyanakkor az ilyen eszközök használata azzal a kockázattal járt, hogy konfliktusba kerülhet a DBA átláthatóságra vonatkozó felelősségével. Ahhoz, hogy a DBA által alkalmazott burkoló módszerek sorát ebben az összefüggésben helyezzük el, először is elemezzük a DBA-nak az ügynökség működésében alkalmazandó mesterséges intelligencia alapú rendszerekkel szemben támasztott követelményekkel kapcsolatos álláspontját. Ez megteremti a terepet azon burkoló módszerek megvitatásához, amelyeket a DBA az ML-megoldások fejlesztése által bevezetett magyarázhatóság-pontosság kompromisszum (lásd az 1. ábrát) kihívásainak kezelésére fejlesztett ki.

4.1 A mesterséges intelligenciára vonatkozó követelmények a DBA-nál

Interjúink azt mutatták, hogy mivel a DBA-nak a mesterséges intelligenciamodellek alkalmazásával javítani kell a működését, jelentős figyelmet kell fordítania arra, hogy az instrumentális eredmények ne járjanak együtt a humanista eredmények figyelmen kívül hagyásával. Két tényező alakította a szervezetet a magyarázhatóság-pontosság kompromisszum egyensúlyának megtalálására irányuló törekvésében: a közhivatalként betöltött pozíciói és az érdekelt felek különböző igényei.

Először is, a DBA-nak mint közhivatalnak jelentős felelőssége van abban, hogy döntései a lehető legigazságosabbak és elfogulatlanabbak legyenek. A közelmúltban a GDPR-hez hasonló szabályozásokkal kapcsolatos viták további figyelmet irányítottak a személyes adatok kezelésére és a polgárok magyarázattal kapcsolatos jogaira. Ezek az okok arra késztették a DBA-t, hogy megbizonyosodjon arról, hogy a szervezet ML-megoldásai kellőképpen megfelelnek a magyarázhatósági követelményeknek. A DBA éves nyilatkozatokkal foglalkozó csapatának egyik vezető tanácsadója, Mary megjegyzése az átláthatóság fontosságával foglalkozik:

> *Úgy gondolom, hogy Dániában általában nagy a bizalom a rendszerek iránt .... Nagyon szeretem az átláthatóságot. Szerintem az a helyes út, ha teljes mértékben nyilvánosságra hozzák, hogy egy rendszer miért reagál [úgy], ahogyan reagál. Ellenkező esetben nem érzi magát biztonságban, hogy a rendszer miért hozza meg azokat a döntéseket, amelyeket meghoz... Számomra nagyon fontos, hogy ne egy fekete doboz legyen.*

Mégis, a DBA-nak bőséges lehetőségei vannak arra, hogy profitáljon a mesterséges intelligencia működésében való alkalmazásából, mivel hatalmas mennyiségű adathoz fér hozzá, és proaktív ügyintézőkkel büszkélkedhet, akik képesek azonosítani a releváns feladatokat a mesterséges intelligencia számára. Néha a megmagyarázhatatlan modellek egyértelműen felülmúlják a megmagyarázható modelleket, így az ügynökséget erősen ösztönzi, hogy a nagyobb pontosság és a jobb teljesítmény elérése érdekében keresse a műveletei számára megvalósítható mesterséges intelligencia-modellek körének bővítésére irányuló lehetőségeket. Ezt azonban úgy kell megtennie, hogy ne vállaljon túlzott kockázatot a megfejthetetlen modellekkel kapcsolatban:

> *Ha az algoritmus kimenete nagyon rossz a [megmagyarázható] modellek használata esetén, és a fejlettebb vagy black-box algoritmusoknál teljesítménynövekedést látunk, akkor [a fejlettebbeket] fogjuk használni. Ezután ellenőrizzük, hogy "oké, hogyan lehet ezt átláthatóvá tenni, hogyan lehet ezt megmagyarázhatóvá tenni..."* (Steven, ML Lab).

Másodszor, a megmagyarázható mesterséges intelligenciára való törekvést még összetettebbé teszi a különböző DBA-érdekcsoportok magyarázattal kapcsolatos követelményeinek sokfélesége. A belső érdekeltek több különböző munkavállalói kategóriát foglalnak magukban, beleértve a vezetőket, az adattudósokat, a rendszerfejlesztőket és az ügyintézőket. Külsőleg a DBA kölcsönhatásban áll a polgárokkal és a Dániában bejegyzett vállalatokkal, valamint a

IT-tanácsadó cégek, amelyek karbantartják az ügynökségnek a termelési környezetben telepített mesterséges intelligenciamodelljeit.

Ezen érdekeltek mindegyike egy adott modell belső logikájának és kimeneteinek sajátos magyarázatát igényli. Míg egy szakértő hasznosnak tarthatja a modell viselkedése mögött meghúzódó logika egy bizonyos fajta magyarázatát, addig ez a magyarázat haszontalan lehet egy olyan személy számára, aki nem szakértő felhasználó. Egy nem szakértő felhasználó számára egy tömör, irányított és akár részben átláthatatlan magyarázat nagyobb értéket képviselhet, mint egy pontos technikai leírás. David, az Early Warning Europe esetfelelőse egy példát hozott fel: "Amikor [egy adattudós] elmagyarázta ezt nekünk, természetesen olyan volt, mint amikor a tanár elmagyarázza ... az agyműtétet egy csoport ötévesnek".

Ez a két tényező együttesen magyarázza, hogy a modelljelöltek körének bővítése miért okozhat problémákat még akkor is, ha pontosabb modellek állnak rendelkezésre és technikailag bevethetők. A különböző érdekelt felek különböző igényei miatt nehéz elérni a magyarázhatóság megfelelő szintjét. Ezért égető szükség van olyan megközelítésekre, amelyek kiszélesíthetik a modellek körét - az 1. ábrán szaggatott körvonallal ábrázolt körként -, és amelyekkel bővíthető a modellek köre.

Eredményeink azt mutatják, hogy a borítékolás lehetséges megoldást kínál a magyarázhatóság és a pontosság közötti kompromisszumra. A különböző borítékolási módszerekkel a megfejthetetlen mesterséges intelligencia kockázatai a különböző érdekeltek számára elfogadható módon ellenőrizhetők, még akkor is, ha a technikai magyarázatok nem állnak rendelkezésre. Ahogy Steven mondta:

> *Gyakran képesek vagyunk [vagyunk] arra, hogy szükség esetén kibontjuk a fekete dobozt, és olyan módon bontjuk ki, hogy az esetkezelőink számára több mint elég jó legyen ahhoz, hogy megértsék és használják, és ahhoz is, hogy elmagyarázzuk, hogyan jutott a modell arra a döntésre, amire jutott.*

Ezután azt tárgyaljuk, hogy a DBA-nak hogyan sikerült ezt elérnie azáltal, hogy a mesterséges intelligencia-rendszerek határait, a képzési adatokat, valamint a bemeneti és kimeneti adatokat burkolta. Ezután megvizsgáljuk megállapításainkat az AI-modell megválasztása és a burkolás közötti kapcsolat tekintetében.

4.2 Boundary Envelopment

A határolás fogalma azt sugallja, hogy egy mesterséges intelligencia-ügynök határait jól meghatározott elvekkel lehet behatárolni, amelyek behatárolják azt a környezetet, amelyen belül feldolgozhatja az adatokat és döntéseket hozhat. A DBA-nál a határok behatárolásának egyik példája a Signature projektben megvalósított dokumentumszűrő. Ez kiszűri azokat a képeket, amelyek nem egy papírdokumentum fotói. Az ilyen szűrő szükségességét akkor állapították meg, amikor egy külső értékelő egy fából készült játékállat képével tesztelte a modellt, és a modell a képet aláírt dokumentumnak ítélte, mivel az

a tervezett környezetén kívül működött. Mivel a modellt nem képezték ki a fekete-fehér dokumentumok szkennelésén és fényképezésén kívül más képek elemzésére, a modell kiszámíthatatlan válaszokat adott. Azáltal, hogy a bemeneti képek típusait azokra korlátozták, amelyek felismerésére a modellt betanították, a válaszként létrehozott szűrő egy olyan határértékként működik, amely garantálja a szükséges változatosságot az információfeldolgozó csővezeték következő elemét alkotó mesterséges intelligencia modell számára. Így a mesterséges intelligencia modell kétféleképpen került burkolásra: technikailag a bemeneti adatok szűrőjének kifejlesztése révén, társadalmilag pedig a munkafolyamat megváltoztatása révén, amelynek során a dokumentumokat mostantól átvizsgálják, mielőtt a teljesség szempontjából értékelnék őket.

A borítékolás társadalmi és technikai dimenziója más esetekben is nyilvánvaló volt az eseti szervezetnél. A következő idézetek azt példázzák, hogy a DBA hogyan hangszereli a mesterséges intelligencia-ügynökök határteremtő munkáját, és hogyan gondoskodik arról, hogy a mesterséges intelligencia-megoldásai a legkülönbözőbb érdekelt felek aggodalmait is megszólítsák. Annak biztosítása érdekében, hogy a mesterséges intelligencia-rendszerek képességei és korlátai ellenőrzöttek és ezáltal körülhatároltak legyenek, a DBA úgy döntött, hogy a mesterséges intelligencia fejlesztését fokozatos szakaszokra osztja, több kis léptékű megoldás bevezetésével, amelyek mindegyike viszonylag egyszerű és jól meghatározott tevékenységek egy bizonyos csoportjára irányul. A következő megjegyzés ezt a módszert foglalja össze:

> *Nos, én egy olyan szervezetnél dolgozom, ahol szerencsére a vezetőség azt akarja, hogy gyorsan fejlesszünk eredményeket, vagy gyorsan bukjunk el, így inkább örülnek annak, ha kis megoldásokat állítanak üzembe [használnak], mintsem hogy nagy projektek bukjanak el .... Úgy döntöttünk, hogy eseményvezérelt architektúrát használunk, mert amikor komplex rendszerekkel foglalkozunk, jobb megengedni a rendezett káoszt, mint megpróbálni a kaotikus rendet. Az eseményvezérelt architektúrával lazán összekapcsolt rendszerekre támaszkodhatunk, és a megbízható metaadatok segítségével rendet teremthetünk az ugyanazon adatokkal kölcsönhatásba lépő különböző rendszerek káoszában.* (Jason, ML Lab)

Így tisztán technikai szempontból az eseményvezérelt architektúra és a lazán kapcsolt rendszerek olyan technikát jelentenek, amelyben egy nagyobb architektúra különböző összetevői autonóm módon működnek, és a meghibásodások csak helyi hatásokra korlátozódnak. Például a hibás döntések kisebb valószínűséggel adódnak tovább más rendszerekre, és ha ez mégis megtörténik, a laza csatolás lehetővé teszi a DBA számára, hogy gyorsan megfékezze a hiba terjedését. Minden egyes komponens tehát a saját burkában működik, és nagyobb burkokat hoznak létre az AI komponensek hálózatként való működésének ellenőrzésére.

Amint azonban a különböző érdekelt felek igényeinek megfelelő borítékokra való fenti hivatkozás is kiemeli,

A határoló borítékok nem csupán technikai célt szolgálnak. Az adatokból vett következő részlet azt mutatja, hogy e határok megértése mennyire fontos azon emberi érdekeltek számára, akiknek feladata a modell működésének helyességének megítélése, amikor például a környezet összetettsége meghaladja a modell felfogóképességét:

> *Körülbelül 160 szabályunk van. Vannak technikai szabályaink, amelyek azt vizsgálják, hogy a megfelelő taxonómiát használják-e, hogy az XBRL formátumú-e, és hogy megfelel-e a követelményeknek. Vannak üzleti szabályaink is. Például, hogy az eszközök és a kötelezettségek egyeznek-e? Néhány szabály csak a technikai kérdéseket vizsgálja a példányjelentésben. Más szabályok olyanok, amelyeket teljes leállításra vonatkozó szabályoknak nevezünk... a benyújtók nem nyújthatják be a jelentést, amíg ki nem javították a hibát. Vannak több iránymutató jellegű szabályaink is, amelyeknél azt mondjuk: "Úgy tűnik, hogy hibát készül elkövetni. A legtöbb ember így csinálja. Biztos, hogy folytatni akarja a jelentés benyújtását?". És akkor [a felhasználók] eldönthetik, hogy figyelmen kívül hagyják-e a szabályt [vagy sem]*. (Mary)

A többféle hiba elszámolásával kapcsolatos technikai kérdéseken túlmenően a megjegyzés a határkeretek társadalmi dimenziójáról is tanúskodik. A határokat világosan elmagyarázzák a DBA belső felhasználóinak, akik szükség esetén felülbírálhatják a modelleket. Ezen túlmenően, az ügyfelekkel szemben álló modellek olyan környezetben működnek, ahol világosan meghatározott szabályok korlátozzák a működésüket. Ahol a nem szakértő alkalmazottak közvetlenül kapcsolatba lépnek egy modellel, ezeket a szabályokat elmagyarázzák nekik, és az embernek mindig lehetősége van arra, hogy figyelmen kívül hagyja a modellek ajánlásait, ha azok megkérdőjelezhetőnek tűnnek.

Ezért fontos, hogy a DBA-nál minden ügyféllel szembenéző mesterséges intelligencia modell esetében a végső határkeret egy ember. Az AI-modell által javasolt döntést mindig egy ügyintéző ellenőrzi. Egyszerűen fogalmazva, az emberi racionalitás létrehoz egy határt, amely körülveszi a modell működését. Ez kettős célt szolgál: megtagadja bármely modell hatalmát arra, hogy felügyelet nélkül hozzon döntéseket, ugyanakkor biztosítja, hogy a DBA minden döntése megfeleljen a jogi követelményeknek. Jason szerint:

> *Az ügynökséget bíróság elé lehet vinni, amikor feloszlatunk egy céget, amikor a törvény segítségével [erőszakkal] megszüntetünk egy céget. És nekünk ebben a helyzetben a bíróságon ... teljes dokumentációt kell benyújtanunk arról, hogy miért hoztuk meg ezt a döntést. Nos, jogi értelemben, amint egy ember is érintett, és ez mindig így van, mindig tartunk egy embert [a] hurokban, [hogy biztosra menjünk]. Ebben a kontextusban, ez csak jogilag*

> *szükséges az emberi döntés bemutatásához. De a döntéstámogatást is meg akarjuk tudni magyarázni, ezért van szükségünk a modellünk és az információs láncunk magyarázhatóságára. A magyarázhatóság a mikroszinten hasznos a szervezet megértéséhez [a] szervezet egyfajta makroszinten.*

Más esetekben a szakértő esetfelelősök állíthatják be a szóban forgó modell küszöbértékeit, hogy az a lehető leghasznosabb és legpontosabb ajánlásokat adja ki. Ennek van egy olyan hatása, amely megkönnyíti a DBA-munkavállalók számára az adott modell elfogadását:

> *Néhány modell esetében [az] általunk meghatározott irányadó küszöbértéket állapítanánk meg. Aztán az esetkezelők szabadon mozgathatják felfelé és lefelé.* (Susan, ML Lab)

> *A modell "elnémításának" vagy a küszöbérték megváltoztatásának lehetősége jelentős kulturális tényező volt e technológia üzleti adaptációjában.* (Jason)

Összefoglalva, a határok behatárolása magában foglalja mind a technikai kérdések megoldását (a modell képességeinek határainak megértése stb.), mind a társadalmi tényezők kezelését (a különböző érdekeltek számára kellő magyarázatot nyújtani, és ezáltal bizalmat kelteni a modell pontossága iránt stb.)

4.3 Képzés-adatok kibontása

A mesterséges intelligencia rendszerek képzéséhez használt adatok döntő fontossága széles körben elismert a mesterséges intelligencia/ML közösségben. Ha különböző adatkészleteken képzik ki, két, egyébként azonos felépítésű modell nagymértékben eltérő kimeneteket produkál (Alpaydin, 2020; Robbins, 2020). Ennek megfelelően a képzési adatok és a képzési folyamat szoros ellenőrzése a borítékolás fontos szempontját képezi: ha a képzési adatok által reprezentált jelenségek spektrumát gondosan figyelembe vesszük, jobban megérthetjük, hogy a modell mit fog tudni - és mit nem - értelmezni.

Mivel a DBA el akarja kerülni, hogy a potenciálisan torzított képzési adatok tengerén szabadon barangoló, ellenőrizetlen modell nemkívánatos eredményeket hozzon, a szervezet úgy döntött, hogy teljes ellenőrzést tart fenn a tanulási folyamat felett; ezért tartózkodik az online tanuló modellek használatától, amelyek önállóan tanulnak tovább a beérkező adatokból. Ez segíti a DBA-t abban, hogy megvédje rendszereit a nem szándékos túlillesztéstől és torzítástól, amelyet a kevésbé szigorúan ellenőrzött képzési adatok könnyebben bevezethetnének. A képzést ellenőrzött, lépcsőzetes módon lehet végrehajtani:

> *Tudatosan döntöttünk úgy, hogy nem használunk [online tanulási] technológiákat, ami azt jelenti, hogy egy bizonyos szintig betanítunk egy modellt, majd elfogadjuk, hogy az nem fog működni.*

> *nem lesz okos, amíg át nem képezzük.* (Jason, ML Lab)

A "menet közben" tanuló modellek elkerülése azzal a hátránnyal jár, hogy a DBA-nál a modellek képzése egy igen bonyolult, emberi szakértelmet igénylő, periodikus folyamat. A sikeres képzési-adatborítékolás ezért azt jelenti, hogy az ügynökségnél számos érdekelt fél rendszeresen együttműködik az átképzési igények felmérése és az átképzés elvégzése érdekében. A képzési adatokra való odafigyelés ösztönzi a belső vitát az adatok alkalmasságáról és a kézi feldolgozásra kijelölt problémás esetek felderítésének lehetséges javításáról.

Az átképzés megfelelő megtervezése érdekében az ML Lab adattudósai rendszeresen kommunikálnak az ügyintézőkkel a modellek teljesítményének és a beérkező új típusú adatok elemzéséről. Bár ez a folyamat időigényes, de támogatja a dolgozók kölcsönös megértését annak, hogy a modellek hogyan jutnak konkrét eredményekre. Egy esetmunkás a következőképpen írta le a hatást:

> *Én nem vagyok olyan technikailag [földhözragadt] ember, de ez - a modell kiképzése és annak megnézése, hogy milyen kimenet jött ki abból, hogy kiképeztem a modellt... - sokkal jobban megértettem a dolgot.* (William, cégbejegyzés)

Az átképzési lépések során megvalósuló interakció révén az érdekeltek jobban megbecsülik egymás igényeit:

> *A vállalati csapatban nagyon szeretnénk [egy olyan modellt, amely] azt mondja nekünk: "Nézd meg ezeket a területeket", olyan területeket, amelyekre nem is gondoltunk: "Nézd meg ezeket, mert látjuk, hogy itt valami romlott dolog folyik", alapvetően. Más ellenőrzési osztályok inkább azt mondanák: "Láttunk egy esetet, amely így nézett ki; ez a nyolc dolog volt rossz. Kedves gép, találj nekem olyan eseteket, amelyek pontosan ugyanilyenek". És mi már sokszor próbáltuk elmondani nekik, hogy ez így rendben van. Évekkel ezelőtt volt egy ügyünk, ahol sok pékség volt, amelyek sok csalást követtek el, de most már nincs értelme pékségeket keresni, mert most ezek a pékségek ... virágokat árulnak, vagy számítógépeket készítenek, vagy valami mást.* (Daniel, cégbejegyzés)

Összefoglalva, a képzési adatok burokba terelése társadalmi erőfeszítéseket foglal magában, a tisztán technikai erőfeszítéssel együtt, amely a megfelelő bemeneti-kimeneti leképezések gépileg olvasható formában történő elkészítésére irányul, hogy aztán a mesterséges intelligencia megtanulhassa azokat. Ahhoz, hogy a képzés-adatburok sikeres legyen, a modell teljesítményének szűrése és folyamatos nyomon követése számos különböző érdekelt fél együttműködését igényli. Csak ez garantálhatja, hogy a torzítások és

az adatok egyéb hiányosságai csökkennek - és a modell naprakész marad. Ellenkező esetben, ahogy a környezet változik a modell körül, a modell határai elavulnak. A képzési adatok burkolása segít ennek kezelésében a torzítással kapcsolatos kérdések mellett.

4.4 Input és Output Envelopment

A bemenet és a kimenet határozza meg, hogy milyen adatforrásokat használnak fel az előrejelzések létrehozásához, és milyen típusú döntéseket, osztályozásokat vagy intézkedéseket hoz létre a modell kimenetként. A jelentős zajt, torzítás kockázatát, adathiányt vagy egyéb problémákat mutató potenciális bemeneteket és kimeneteket ezeken a döntéseken keresztül kizárják a mesterséges intelligencia működéséből. A bemeneti források kiválasztása tehát szorosan kapcsolódik az adatminőséggel kapcsolatos elképzelésekhez. A PassportEye személyazonosság-felismerő modell konkrét esetében a laboratórium munkatársai számára világossá váltak a bemeneti ellenőrzés előnyei a gyenge és változó végfelhasználó által generált tartalom körülményei között:

> *Azt hiszem, a fő problémánk az volt, hogy igen, egy kicsit oda-vissza kellett mennünk, mert a bemeneti adatok nagyon eltérő minőségűek voltak. Többnyire alacsony minőségűek voltak. A PassportEye a dobozból kivéve valójában nagyon rossz eredményeket adott vissza, és ez tükrözi a bemeneti adatok alacsony minőségét, mivel az emberek bármilyen megvilágításban, [bármilyen háttérrel szemben] bármilyen képet készítenek, és így tovább. Ezért kitaláltuk, hogyan lehet a képeket előre-hátra forgatni, hogy megbízhatóbb eredményt kapjunk. Kiderült ugyanis, hogy a PassportEye meglehetősen érzékeny a kép szögére. Nem mi írtuk [a képelemző szoftvert], így ez talán az egyik kockázatos része annak, amikor csak importálsz egy könyvtárat ahelyett, hogy magad írnád meg.* (Thomas, ML Lab)

Ami a kibocsátás-összegzést illeti, itt a társadalmi és a technikai kölcsönhatás sokkal hangsúlyosabban jelenik meg. Ahelyett, hogy egyszerűen megakadályozná a megbízhatatlan kimenetek előállítását, a DBA árnyaltabb megközelítést alkalmaz. A modellekből származó megfelelő megbízhatósági osztályzatok és intervallumok kiadása a DBA aktív tanácskozás tárgyát képezi. Az olyan becslések, mint például egy pénzügyi dokumentum aláírásának valószínűsége, fontosak az ügynökség ügyintézői számára, akiknek szükségük van rájuk a problémás esetek azonosításához. Ha egy mesterséges intelligenciamodell egyértelműen meghatározott és érthető konfidenciaértéket ad, akkor az ügyintézők figyelmét szükség esetén gyorsan fel lehet hívni a modell kimenetére:

> *Ha nincs aláírás, [az ügyintézők] egyszerűen elutasítják. Mert a törvény azt mondja, hogy ezt a dokumentumot alá kell írni, tehát az ember ránéz a papírokra, és azt mondja: "Nincs itt. Nem fogja megkapni az áfa-számát, vagy a cégszámát, mert nem írta alá a dokumentumot. "* (James, ML Lab)

Ha az ügyintéző a bizalmi minősítések alapján képes ellenőrizni az ítéleteket, akkor a DBA-ügyfelekkel (pl. a dokumentumokat benyújtó vállalatokkal) való interakciók során felelősségteljesen tud eljárni, és meggyőzően tud válaszolni a megkereséseikre. Ahogy Steven kifejtette:

> *Ha valaki felhívja, és megkérdezi: "Miért utasították el a dokumentumomat?", akkor az ügyintéző azt fogja mondani: "Azért, mert nem írta alá". "Honnan tudja ezt?" "Megnéztem a dokumentumot. Nincs aláírva." Így nem kell azt válaszolniuk, hogy "Nos, a neurális hálózat azt mondta, hogy azért, mert a sarokban van egy 644-es változó". Ezért lehet megúszni [a] neurális hálózat használatát ebben az esetben, függetlenül a megmagyarázhatóságtól.*

Néha azonban nehezebb egyértelműen ellenőrizni a modell kimenetét, ilyenkor a szervezet arra törekszik, hogy a modell kimenetének társadalmi kontextusát ismerő területi szakértőkkel konzultálva megértse a mesterséges intelligencia modell viselkedését. Ahogy Steven fogalmazott: "Amikor [nehezebb] megállapítani, hogy a modell helyes vagy helytelen, akkor az eseteket az ügyintézőkhöz toljuk, és azt mondjuk: "Kérem, nézze meg ezt".".

A bemeneti és kimeneti burkolás e példái a társadalmi és a technikai tényezők közötti egyértelmű kölcsönhatást mutatják. Míg egy átláthatatlan modell képes nagy mennyiségű strukturálatlan adat hatékony feldolgozására és ajánlások készítésére az egyes dokumentumok elfogadására vagy elutasítására vonatkozóan, ezt a folyamatot szorosan irányítják az ügyintézők, akik a szervezeti célkitűzésekre és a jogszabályi korlátozásokra támaszkodva biztosítják, hogy az AI által hozott döntések összhangban legyenek az igényeikkel. A végső döntések tehát az emberek és a mesterséges intelligencia tevékenységeinek metszéspontjában születnek.

4.5 Az Envelopment következményei a modellválasztásra vonatkozóan

Miután bemutattuk több borítékolási módszer együttes használatát a DBA-nál, most rátérünk a megfelelő mesterséges intelligenciamodell kiválasztására gyakorolt hatásukra. Összességében a borítékolási módszerek elfogadása lehetővé tette a DBA számára, hogy olyan modelleket is használjon, amelyek egyébként kockázatot jelenthetnének. A különböző mesterséges intelligencia modellek különböző architektúrákon alapulnak, ami kihatással van arra, hogy a modellek mit tudnak és mit nem. A modellek különböznek például érettségük, zajjal szembeni robusztusságuk, a tanulás megszüntetésének és gyors újratanításának képessége és skálázhatóságuk tekintetében. Ezek a tulajdonságok a modelltípus kiválasztásától függnek. Például a zajjal szembeni robusztusságot gyakran könnyebb elérni a neurális hálózatokkal, míg a gyors visszatanulás és átképzés képességei gyorsabban kihasználhatók a döntési fákkal. Az adott modelltípussal kapcsolatos pontossági és/vagy magyarázhatósági igényektől függően, a felhasználási eset mellett, a megfelelően kiválasztott burkoló módszerek a következőképpen valósíthatók meg

rétegek, amelyek együttesen garantálják a biztonságos és kiszámítható működést.

A Boundary Envelopment több szabadságfokot adott a DBA-nak a modellek kiválasztásában azáltal, hogy korlátozta a mesterséges intelligencia ágens befolyási körét. Ez lehetővé tette a munkatársak számára, hogy olyan összetett modelleket használjanak ki, amelyek a burkolás nélkül a megmagyarázhatóság hiánya miatt problémásak lennének. Jason ezt a következőképpen jellemezte: "Mondhatni, hogy szervezetileg egy-egy kis keksszel etetjük a sárkányt, így olyan modelleket tudunk előállítani, amelyeket be lehet vinni a termelésbe, és valóban be is vezetik őket". Ily módon az emberi ágensek a szervezet folyamatait és struktúráit úgy igazítják ki, hogy a technológiai ágens műveleteit biztonságosan féken tartsák.

Hasonlóképpen, az adatok megértése és ellenőrzése a képzési adatok és a bemeneti adatok burkolásával együtt garantálja, hogy a modell viselkedése biztonságos határokon belül marad, és hogy a DBA kellőképpen megérti, hogyan keletkeznek a kimenetek, még a teljes technikai nyomon követhetőség hiányában is. Ahogy James az ML Lab-nál elmélkedett:

> *Itt egy új adatsor. Mit mondhatunk róla? Mire kell odafigyelnünk? Ez egyre fontosabbá válik, mivel egyre több adatot használunk az emberek egyéni jövedelméhez kapcsolódóan, ami Dániában titkos .... A modell kezdeti használatával kapcsolatos tapasztalataink ... hangsúlyozták, hogy ennek a modellnek és az általa [felölelt] adatoknak további irányításra van szükségük annak biztosítása érdekében, hogy ne lépjünk túl az eredeti szándékainkon ... Újra átnéztünk néhány, a platformba beépített metaadat-kezelést annak biztosítása érdekében, hogy megkapjuk a szükséges adatokat arról, hogy a modell hogyan viselkedik az esetkezeléssel kapcsolatban, hogy felmérhessük a modell kimenetét.*

Ami a kimenetet illeti, feltéve, hogy egy ember képes megítélni annak érvényességét, könnyen választhatjuk a feketedobozos modelleket, amelyek jobb teljesítményt nyújtanak. James következő megjegyzése azt mutatja be, hogy a kimenet ellenőrzésének gyakorlása hogyan tette lehetővé egy kifürkészhetetlen modell használatát: "Nem kell tudnom megmagyarázni, hogyan jutottam el az eredményhez olyan esetekben, mint például egy aláírás azonosítása egy papíron. Egyszerűen csak mélytanulást kell végezni, mert azt utólag könnyen ellenőrizheti egy ember."

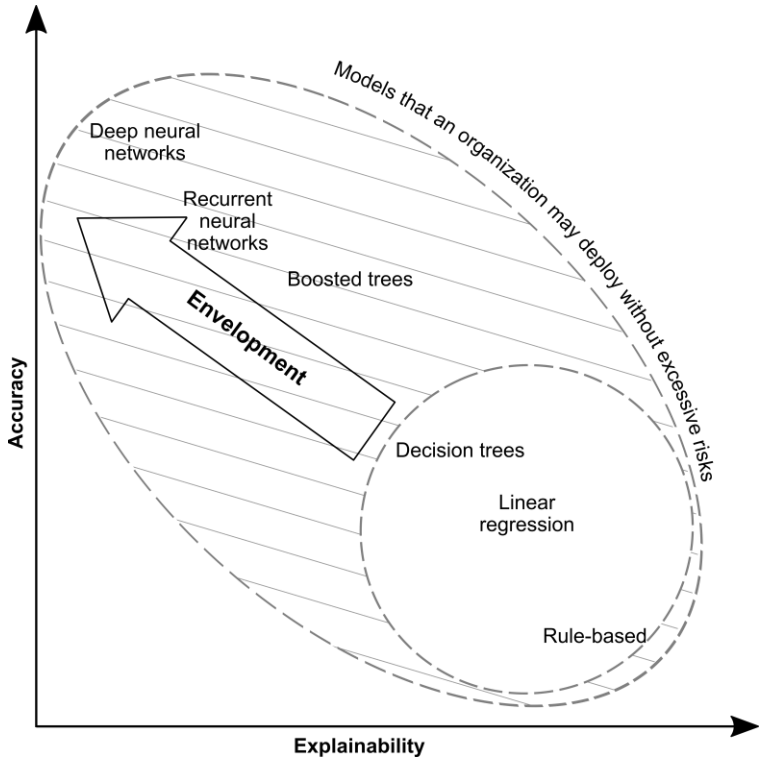
Az interjúk jól szemléltetik, hogy az új modellek iránti igény felmerülhet új jogalkotási kezdeményezések, új szervezeti stratégia vagy az adófizetői magatartás változásai miatt. Előfordulhat, hogy egy meglévő modellt át kell képezni vagy akár teljesen át kell alakítani, ha a pontosság vagy a megmagyarázhatóság mérőszámai azt mutatják, hogy az már nem megfelelő.

kielégítően teljesít (pl. az osztályozásai már nem pontosak, vagy értelmetlen becslésekhez vezetnek, amelyeket nem lehet megmagyarázni). James egy példával illusztrálta, hogy ilyen esetben egy határfelületi burkológörbe segítségével "elnémít" egy modellt, amíg azt átképzésre vagy cserére irányítják: "Az esetfelelősök úgy találták, hogy a modell kimenete nem olyan minőségű, hogy bármihez is használni tudnák, ezért elnémították a modellt. Ez visszahat ránk. Levesszük a modellt. Újratanítjuk...." Ezzel a folyamattal az emberek csökkentették a mesterséges intelligencia ügynöki szerepét a munkafolyamatban azáltal, hogy elnémították, és újratanítással vagy cserével újratárgyalták az ügynöki szerepét.

4.6 Összefoglaló

A borítékolás fogalma segített nekünk abban, hogy konkretizáljuk a kifürkészhetetlen mesterséges intelligencia által támasztott kihívásokkal szembeni fellépés koncepcionális és gyakorlati mechanizmusairól alkotott elképzeléseinket. A fenti alfejezetek empirikus bizonyítékot szolgáltatnak több különböző borítékolási módszerre szervezeti környezetben. Érdemes megjegyezni, hogy míg bizonyítékot találtunk arra, hogy a DBA aktívan alkalmazza a határérték-, a képzési adatok, valamint a bemeneti és kimeneti adatok borítékolását, nem figyeltünk meg vitákat a Robbins (2019) által felsorolt öt borítékolási módszer közül az utolsóról: a funkcióborítékolásról, amely az olvasó talán emlékszik rá, hogy annak eldöntésére utal, hogy egy AI-ügynököt nem használnak bizonyos célokra, még akkor sem, ha azt pontosan meg tudná tenni. E döntés mögött például etikai megfontolások állhatnak. Úgy véljük, hogy a funkcióborítékolással kapcsolatos témák megvitatásának hiánya a DBA-nál azzal magyarázható, hogy az egyes rendszerek céljait már szűken meghatározták az egyes folyamatokra vonatkozó kormányzati előírások alapján.

Az eredményeket a következőképpen foglaljuk össze. Figyelembe véve egyrészt azt, hogy a DBA több mesterséges intelligencia-alapú megoldást is sikeresen alkalmazhatott működésében, másrészt pedig a DBA gyakorlatában (általánosságban és a különböző módszerekre vonatkozóan is) a borítékolásra utaló jeleket, úgy tűnik, hogy a borítékolás fogalma hatékonyan megragad néhány olyan módot, amelyekkel az 1. ábrán bemutatott magyarázhatóság-pontosság kompromisszumot a mesterséges intelligencia megvalósítása során kezelni lehet. Konkrétabban, eredményeink azt mutatják, hogy bár a borítékolás nem változtatja meg a pontosság és a megmagyarázhatóság közötti kapcsolatot, lehetővé teszi a szervezetek számára, hogy az AI-modellek szélesebb köréből válasszanak anélkül, hogy a káros következmények (pl. vadul kiszámíthatatlan eredmények) leküzdhetetlen kockázatával kellene szembenézniük. Az envelopment lehetővé teheti egy szervezet számára, hogy a nagyobb pontosság kedvéért némi magyarázhatóságot kompromittáljon, anélkül, hogy aggódnia kellene, mindaddig, amíg ez a kiszámítható viselkedés bizonyos határain belül történik. A borítékolás legfőbb előnyét az alábbi 3. ábra mutatja be.



3. ábra. Hogyan bővíti az Envelopment a modellek körét, amelyeket egy szervezet túlzott kockázatok nélkül elfogadhat?

Másodszor, a szociotechnikai perspektívát tekintve, függetlenül attól, hogy melyik borítékolási módszerről beszéltek, a megkérdezettek soha nem beszéltek pusztán technikai megoldásról az AI-ügynökök képességeinek korlátozására. Az elemzés feltárta, hogy az ilyen intézkedések nem elszigetelten, hanem mindig iteratív tárgyalásokon keresztül történtek, amelyek figyelembe vették az érdekelt felek többféle nézetét, a társadalom iránti felelősséget és a személyzet munkafolyamataira gyakorolt konkrét következményeket.

5 Megbeszélés

Ebben a kutatásban azt a kérdést tettük fel: *Hogyan tudja egy szervezet biztonságosan és társadalmilag felelős módon kihasználni a kiismerhetetlen mesterséges intelligencia rendszereket*? Erre a kérdésre kerestük a választ egy olyan közpénzből finanszírozott szervezet esettanulmányával, amely rendszeresen alkalmaz mesterséges intelligenciát a társadalom számára fontos működésének javítása érdekében. A fent leírtak szerint a tanulmány és az eredmények elemzése a borítékolás koncepciójára épített, mint a pontosság és a megmagyarázhatóság közötti egyensúly megteremtésének, valamint a hatékonyság és a biztonság közötti jó összhang megtalálásának lehetséges megközelítésére.

A fent bemutatott elemzés egyértelműen három jelentős megállapítást tett. Először is, az esettanulmány megmutatta, hogy az AI borítékolása, mint fogalom, empirikus érvényességgel bír egy szervezeti tudás-munka környezetben. Ez kiegészíti a korábbi borítékolási szakirodalmat (lásd Floridi, 2011; Robbins, 2020), amely tisztán koncepcionális jellegű. Másodszor, kimutattuk, hogy a borítékolás sokkal több, mint egy technikai kérdés - ahhoz, hogy hatékony legyen, az kell, hogy legyen

a technikai és a társadalmi kérdések metszéspontjában kell elhelyezkednie. Tanulmányunk megmutatta, hogy a társadalmi tényezők hogyan hatják át a burkolás minden aspektusát, és hogy az emberi szereplők a burkolás szerves részét képezik, akik felelősek a megfelelő burkolatok meghatározásáért, valamint azok fenntartásáért és újratárgyalásáért. Végül az elemzés összefüggéseket mutatott ki a borítékolási módszerek és az ML-modell kiválasztása között. Ezek az eredmények együttesen bizonyítják a borítékolás - különösen a szociotechnikai borítékolás - hasznosságát, mint olyan megközelítés, amely lehetővé teszi annak megértését, hogy miként fogalmazható meg a mesterséges intelligencia szerepe egy szervezetben, és miként határozhatók meg és irányíthatók a felelősségi körök. A következőkben az elméletre és a gyakorlatra vonatkozó konkrét következményekkel foglalkozunk.

5.1 Következtetések az elméletre

A fent leírt megfontolások figyelembevétele lehetővé teszi a burkolt mesterséges intelligencia mélyebb szociotechnikai megvitatását, a DBA példáján keresztül. Ez a korábbi szakirodalom és az empirikus eredményeink szintézise révén lehetséges. Sarker et al. (2019) áttekintése az IS-kutatás szociotechnikai megközelítéseiről, amelyet e tanulmány elején tárgyaltunk, arra figyelmeztet, hogy a mai IS-munkát az a veszély fenyegeti, hogy túl gyakran a technológiák instrumentális eredményeire összpontosít, mivel azokat könnyebb mérni és értékelni. Sarker és munkatársai azt javasolják, hogy a szociotechnikai irányultságú IS-kutatók jól tennék, ha a rendszerek instrumentális és humanista eredményeivel egyaránt foglalkoznának.

A DBA esetében a mesterséges intelligencia lehetséges instrumentális eredményeit valóban könnyebb lenne elemezni és kijelenteni, mint a humánus eredményeket, mivel ezek a folyamatok automatizálásának tipikus okaihoz kapcsolódnak, mint például a hatékonyság növelésének és a nagyobb pontosságnak a céljai. Láttuk azonban, hogy a DBA nem csak az ilyen instrumentális eredményeket veszi figyelembe: döntő fontosságúnak tartották, hogy a mesterséges intelligencia projektek ne vezessenek a kormányzati hatalommal való visszaéléshez vagy a polgárok vagy a magánvállalkozások szükségtelen profilozásához/megfigyeléséhez. Az ilyen eredmények humanista szempontból problematikusak lennének, és veszélyeztetnék a szervezet mint hatóság integritását, ami a közbizalom erózióját eredményezhetné. A mesterséges intelligencia projekteknek ráadásul a DBA-n belül is vannak humanista eredményei. Bővítik az ügyintézők lehetőségeit a munkafolyamatok újratervezésére - valójában az ügynökség legtöbb projektjét az ő javaslataik alapján indítják -, és az ügyintézők közvetlenül részt vesznek az AI fejlesztési folyamatokban is. Ez a munkahelyi demokrácia, a felhatalmazás és a munkahelyi jólét növelését szolgálja. A DBA mesterséges intelligenciával való körülbástyázása egyértelműen szociotechnikai folyamat: a mesterséges intelligencia működésére vonatkozó korlátok technikai meghatározása egy olyan társadalmi folyamaton keresztül történik, amelyben az esetmunkások és más érdekeltek központi szereplők.

Az a tény, hogy a DBA mesterséges intelligencia fejlesztését jellemzően az ügyintézők indítják el, arra utal, hogy a szervezet emergens működési módot fogadott el. Az esetmunkások azonosítják a gyakorlati területi problémákat, amelyeken az ML Lab dolgozik, és részt vesznek a mesterséges intelligencia modellek fejlesztésében is. A megfelelő modell keresése során az ML-szakértők és az eseti dolgozók elemzik a különböző ML-modellekkel járó képességeket és korlátozásokat, majd interaktív módon összevetik azokat a megoldandó problémák tulajdonságaival. Ha az adott problémához nem találnak megfelelő modelleket, a problémát alternatív struktúrára bontják. Egy másik megközelítés ilyen esetekben az, hogy az esetmunkások megoldásban betöltött szerepét úgy alakítják át, hogy az illeszkedjen a mesterséges intelligencia rendszer képességeihez.

Elméleti implikációkat javasolunk (1) a szervezeti mesterséges intelligencia megvalósításának az emberi és a mesterséges intelligencia közötti egyensúlyozásként való leírására, és (2) a szociotechnikai burkolásnak mint e kulcsfontosságú egyensúlyozás elsődleges eszközének konceptualizálására. Az első implikáció kezelése arra épül, hogy az AI fejlesztési folyamatok olyan cselekvéssorozatokból állnak, amelyekben az esetmunkások és az AI rendszerek, mint partneri ágensek, együttesen hajtanak végre feladatokat. Az ügynöki tevékenység kívánt szintjét (azaz az emberek és a mesterséges intelligencia rendszerek közötti megfelelő egyensúlyt) a modellek fejlesztése során határozzák meg, és a lehetséges mesterséges intelligencia megoldások képességei és korlátai határozzák meg. A mesterséges intelligencia technológiák nagy teljesítményű információfeldolgozó

képességek rengeteg lehetőséget kínálnak a végrehajtás számos fajtájához (Kaplan & Haenlein, 2019). Ugyanakkor a skálázható számítási erőforrások könnyű elérhetőségének köszönhetően az AI kevés korlátot szab az adatfeldolgozási kapacitásnak (Lindebaum et al., 2020). Ezért az ilyen technológia felhasználására számtalan lehetőség kínálkozik. Számos AI-modell összetettsége miatt azonban a technológia korlátokat támaszt a működésének technikai magyarázatára való képességét illetően. Ezért a mesterséges intelligenciában rejlő lehetőségeket még mindig megfelelően korlátozni kell: például meg kell találni egy elfogadható magyarázhatóság-pontosság kompromisszumot, és ehhez meg kell határozni az adott kontextusban az értelmes magyarázhatóság szükséges szintjét is (Ribera & Lapedriza, 2019; Robbins, 2019), ami a társadalmi szereplők közötti, ügynökségen átívelő tárgyalásokon keresztül történik. Ezért a mesterséges intelligencia megvalósításai általában az emberi és a mesterséges intelligencia ügynöksége közötti egyensúlyozással járnak, hogy a mesterséges intelligencia számára megfelelő szintű ügynöki szintre jussanak. Ebben az összefüggésben a két fél közötti hatalmi egyensúly kiegyenlítettebb, mint sok más ember-technológia viszonyban (pl. vállalati erőforrás-tervezési rendszerek bevezetése), ahol a technológia működése ismert, és képességei kevésbé valószínűnek tűnnek, hogy váratlan negatív következményeket jelentenek az érdekeltek számára.

Ez a vita elvezet bennünket a második implikációhoz: a szociotechnikai burkolás konceptualizációjához. Az ilyen jellegű kétirányú borítékolás hangsúlyozza a társadalmi dimenziót, amely hiányzik a meglévő borítékolási irodalomból (Floridi, 2011; Robbins, 2020), azáltal, hogy az emberi és az AI-ügynökségek kölcsönhatására összpontosít, ahelyett, hogy pusztán az AI-rendszer képességeinek korlátozására vagy beállítására összpontosítana. Ezzel sikerült kiterjeszteni a borítékolásról szóló vitát azáltal, hogy feltárjuk, hogyan lehet a borítékokat szociotechnikai környezetben felépíteni és fenntartani. Úgy véljük, hogy a burkolásnak ez a szociotechnikai szemlélete hatékony eszközt nyújthat az emberi és az AI-ügynökség közötti egyensúlyozás sikeréhez, mivel olyan gazdag mechanizmust kínál, amelyen keresztül az AI-képességek korlátozhatók olyan környezetben, ahol az etika, a biztonság és az elszámoltathatóság létfontosságú a műveletekhez. Ez segíthet ellensúlyozni a mesterséges intelligencia megfejthetetlen volta által bevezetett bizonytalanság hatását, és így lehetővé teszi a szervezetek számára, hogy hatékonyságnövekedést érjenek el a nagy teljesítményű, de megmagyarázhatatlan képességeket kínáló mesterséges intelligenciarendszerekből.

5.2 Gyakorlati következmények

A menedzserek számára, akik gyakran inkább az emberek, mint a mesterséges intelligencia-ügynökök irányításában jártasak, az ebben a tanulmányban bemutatott és illusztrált burkoló módszerek megfelelő szókincset és eszköztárat kínálnak a mesterséges intelligencia fejlesztésének kezeléséhez.[1] Az adott mesterséges intelligencia kockázatainak elemzése révén egy adott AI

[1] A DBA esete alapján készült részletesebb vezetői ajánlásokért lásd Asatiani et al. (2020).

megoldást hoz létre az üzleti élet, az etika, a fogyasztói jogok (pl. a magyarázathoz való jog) és a környezetbiztonság számára, a vezető képes lehet felfogni a szervezet borítékolási igényeit. Ennek alapján a szociotechnikai megközelítések megvalósíthatók és összehangolhatók a műveletirányítással és az AI-megoldások fejlesztésével, mindezt oly módon, hogy a modellek érthetőbbé válnak az érdekeltek számára, és az adattudósokra jellemző AI értelmezhetőségi igényeket kezelik.

Az óvatosság azonban elengedhetetlen. Még burkolás jelenlétében sem szabad elfogadni a fekete doboz modelleket anélkül, hogy jelentős erőfeszítéseket tettünk volna értelmezhető modellek megtalálására. Bár kezdetben úgy tűnhet, hogy a fekete doboz modell az egyetlen alternatíva, jó okunk van azt hinni, hogy a jelenleg felismertnél sokkal több területen létezhetnek pontos, de értelmezhető modellek. Az ilyen modellek azonosítása nagyobb előnyt kínál, mint a fekete dobozos modellek szociotechnikai burkolásának. Minden bizonytalansággal és korlátozott gyakorló adathalmazzal járó döntési problémára általában számos közel optimális, ésszerűen pontos előrejelző modell azonosítható. Ez az állítás az úgynevezett Rashomon-halmaz érvből ered (Rudin, 2019), amely szerint jó esély van arra, hogy az elfogadható modellek közül legalább egy értelmezhető, mégis pontos. Egy másik ajánlott megközelítés, amely egyszerűsíti a borítékolást, a "szürke dobozos modellekre" való törekvés, ahogyan azt a valós, fizikai folyamatokat szimulálni képes "digitális ikrek" létrehozása példázza (lásd El Saddik, 2018; Kritzinger et al., 2018). A szürke dobozos ML-megoldásokat az adott területen ismert törvények, elméletek és elvek szerint modellezik. Egy ilyen megközelítéssel például létrehozható egy neurális hálózat struktúrája, amely után a szabad paraméterek gyorsabban betaníthatók a nagy teljesítmény elérése érdekében, a magyarázhatóság csökkenése nélkül.

A borítékolásnak mint az AI megvalósításának eszközének elfogadásából származó másik gyakorlati előnye a technikai adóssághoz való viszonya. A mesterséges intelligencia kontextusában legalább kétféle adósságot lehet azonosítani. Az első az olyan modellek kiválasztásához kapcsolódik, amelyek nem a legjobb pontosságot kínálják az adott problémákhoz (Cunningham, 1992; Kruchten et al,. 2012), ami akkor fordul elő, ha egy szervezetnek magyarázhatóságot kell biztosítania a megvalósítás során. A másik, a dokumentációval kapcsolatos forrás általában a szoftverfejlesztésre vonatkozik: a szervezetek dönthetnek úgy, hogy felgyorsítják a megvalósítási erőfeszítéseiket, ha úgy döntenek, hogy lazítanak a döntéseik és a kód dokumentálására vonatkozó követelményeken (lásd Allman, 2012; Rolland et al., 2018). Ez visszafelé sülhet el, ha a dolgozói fluktuáció felüti a fejét, és nem marad senki, aki el tudná magyarázni a mesterséges intelligencia rendszer mögöttes logikáját. Végül is a válaszok csak az egyének fejében léteznek, vagy a kódban elásva.

Az envelopment talán mindkét típusú adósság kezelésére alkalmas: a mesterséges intelligencia megvalósítása során hozott kockázatkerülő döntésekből eredő adósság, amely elmarad a probléma megoldásától.

fejlesztés, valamint a dokumentációs követelmények enyhítésére vonatkozó döntések miatt bekövetkező adósságok. Mivel a borítékolás magában foglalja a döntések gondos meghozatalát és dokumentálását, olyan gyakorlatként szolgálhat, amelynek révén a tervezési döntések explicitté válnak; például a problémával és a modellel kapcsolatos implicit feltételezések rögzíthetők. A burkolás tehát nemcsak a dokumentációt támogatja, hanem a pontosabb modellek használatának lehetővé tételével a konzervatív modellválasztási stratégiában gyökerező technikai adósságok felhalmozódását is csökkentheti.

5.3 Korlátozások és további kutatások

Kutatásunknak vannak bizonyos korlátai. Először is, célzott mintavételt alkalmaztunk, és egy kormányzati egységet vizsgáltunk empirikus esetként, mivel feltételeztük, hogy ez empirikusan gazdag környezetet biztosít a mesterséges intelligencia használatára vonatkozó adatok gyűjtéséhez. Ez a választás, bár bőséges bizonyítékot szolgáltatott az alkalmazott burkoló stratégiákról, arra korlátozott minket, hogy az ilyen stratégiákat egy állami szervezet sajátos környezetében vizsgáljuk. A további kutatások a mesterséges intelligencia szélesebb körben vizsgálhatnák a mesterséges intelligenciát. Például a különböző súlyozású célok által vezérelt magánvállalkozások más típusú burkoló stratégiákat alkalmazhatnak, vagy az általunk vizsgált stratégiákat más módon alkalmazhatják. Ezenkívül tanulmányunk nem talált bizonyítékot a funkcióborítékolásra vonatkozóan - valószínűleg azért, mert az AI használatának céljait a DBA-nál már szigorúan törvények és rendeletek írják elő. Valójában ritkán volt okunk megvitatni, hogy a DBA mesterséges intelligencia megoldásait olyan célokra kellene-e alkalmazni, amelyekre azokat soha nem tervezték. Másodszor, bár az ügyviteli szervezethez való hozzáférésünk lehetővé tette az alkalmazott burkoló stratégiák mélyreható elemzését, nem tudtuk megvizsgálni azok hosszú távú következményeit. További kutatásokra van szükség ahhoz, hogy megvizsgáljuk e borítékolási stratégiák időbeli hatásait. Végül, bár nagyvonalú hozzáférést kaptunk az interjúk lefolytatásához és a másodlagos anyagok elemzéséhez, az interjúkból származó adataink természetesen az informátorok által elmondottakra korlátozódnak. Az informátorok elfogultságával kapcsolatos kockázatok csökkentése érdekében arra törekedtünk, hogy a burkoló stratégiákkal kapcsolatos valamennyi kritikus bizonyítékról többféle véleményt kapjunk. Például interjút készítettünk a DBA ML Laboratóriumában dolgozó minden alkalmazottal, azzal a céllal, hogy minden egyes projektre vonatkozóan több nézőpontot is kiaknázzunk.

Mind e dokumentum hasznosságát, mind pedig az itt bemutatott erőfeszítésekből származó eredményeket tekintve hangsúlyozni kívánjuk, hogy a mesterséges intelligencia és az ML megoldások különböző módszereinek teljesebb megértése fontos annak érdekében, hogy kiaknázzuk az általuk kínált erősségeket. A borítékolási stratégiák és azok mélyebb vizsgálata gyakorlati eszközt kínálhat e cél eléréséhez. Bár a borítékolás alkalmazása a DBA-nál nem az ezeket konceptualizáló szakirodalomban volt megalapozva.

gyakorlatok (pl. Floridi, 2011; Robbins, 2020), mivel a DBA fejlesztők ismerik ezt a korábbi munkát, a módszerek potenciáljának megalapozottabb kiaknázása következhet. Az ilyen lehetőségek mellett a jövőbeli kutatások megvizsgálhatnák, hogy az itt tárgyalt, emberek és mesterséges intelligencia-ügynökök közötti dinamika átvihető-e a mesterséges intelligencia megvalósításától eltérő kontextusokra is. Úgy véljük, hogy hasonló logika azonosítható, bár más formában, más olyan kontextusokban, ahol a biztonságos, etikus és elszámoltatható mesterséges intelligencia megvalósítása kulcsfontosságú.

6 Következtetés

A szociotechnikai borítékolás szervezeti kontextusban történő meghatározásában és operacionalizálásában jelentős ígéretet látunk. Az eredmények rávilágítanak a borítékolás konkrét eseteire, és segítenek a konkrét társadalmi és technikai orientált a burkolás megközelítései. Kiindulópontként egy ínycsiklandó pillantást tudtunk nyújtani a különböző szociotechnikai burkolózási megközelítések lehetőségeire és korlátaira a mesterséges intelligencia biztonságosabb, az emberi javak érdekében történő felhasználásával kapcsolatos kérdések kezelésében.

Köszönetnyilvánítás

Hálásak vagyunk a dán üzleti hatóságnak és az Early Warning Europe-nak a tanulmány elkészítésének lehetőségéért. Szeretnénk köszönetet mondani a különszám szerkesztőinek és három névtelen bírálónak, akiknek éleslátó észrevételei és építő jellegű kritikája segítettek nekünk abban, hogy jelentősen javítsuk tanulmányunk minőségét. Köszönjük továbbá az ICIS 2019 JAIS/MISQE különkiadásának kerekasztal-beszélgetésén részt vevőknek a projektjavaslatunkkal kapcsolatos visszajelzéseiket. Természetesen minden fennmaradó hiba a miénk.

Hivatkozások

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Egy emberhez hasonló, nyílt tartományú chatbot felé. https://arxiv.org/pdf/2001.09977v1.pdf.

Ågerfalk, P. J. (2020). A mesterséges intelligencia mint digitális ügynökség. *European Journal of Information Systems*, *29*(1), 1-8.

Allman, E. (2012). A irányítása. műszaki adósság kezelése. *Communications of the ACM, 55*(5), 50-55.

Alpaydin, E. (2020. *Bevezetés a gépi tanulásba*, (4. kiadás). MIT Press.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2020). A fekete dobozos mesterséges intelligencia rendszerek viselkedésének magyarázatával kapcsolatos kihívások. *MIS Quarterly Executive, 19*(4), 259-278.

Asatiani, A., & Penttinen, E. (2019). Folytonosságok konstruálása virtuális munkakörnyezetekben: Egy több esettanulmány két, a virtualitás eltérő fokával rendelkező vállalatról. *Information Systems Journal, 29*(2), 484-513.

Asatiani, A., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2019). Az automatizálás mint elosztott megismerés megvalósítása tudásalapú munkaszervezetekben: Hat ajánlás a vezetők számára. *Proceedings of the 40th International Conference on Information Systems*.

Ashby, W. R. (1958). Szükséges változatosság és ennek következményei a komplex rendszerek irányítására. *Cybernetica, 1*(2), 83-99.

Belgrave, L. L., & Seide, K. (2019). Kódolás a megalapozott elmélethez. In A. Bryant and K. Charmaz (eds.), *The SAGE Handbook of Current Developments in Grounded Theory*, (pp. 167- 185). SAGE.

Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Különszám szerkesztői cikke: Mesterséges intelligencia a szervezetekben: Jelenlegi helyzet és jövőbeli lehetőségek. *MIS Quarterly Executive, 19*(4), ix-xxi.

Benbya, H., & McKelvey, B. (2006). A koevolúciós és komplexitáselméletek használata az IS összehangolásának javítására: A multi-level approach. *Journal of Information Technology, 21*(4), Springer, 284-298.

Bernard, H. R. (2017). *Kutatási módszerek az antropológiában: Minőségi és mennyiségi megközelítések*. Rowman & Littlefield.

Bostrom, R., Gupta, S., & Thomas, D. (2009). Metaelmélet az információs rendszerek megértéséhez a szociotechnikai rendszerekben. *Journal of Management Information Systems, 26*(1) 17-48.

Briggs, R. O., Nunamaker, J. F., & Sprague, R. H. (2010). Különleges szekció: A szociotechnikai rendszerek társadalmi vonatkozásai. *Journal of Management Information Systems, 27*(1), 13-16.

Brynjolfsson, E., & McAfee, A. (2014). *A második gépkorszak: Munka, haladás és jólét a briliáns technológiák korában*. Norton.

Burrell, J. (2016). Hogyan "gondolkodik" a gép: A gépi tanulási algoritmusok átláthatatlanságának megértése. *Big Data and Society, 3*(1), 1-12.

Butler, B. S., & Gray, P. H. (2006). Megbízhatóság, tudatosság és információs rendszerek. *MIS Quarterly, 30*(2), 211-224.

Charmaz, K. (2006). *A megalapozott elmélet konstruálása: Gyakorlati útmutató a kvalitatív elemzésen keresztül*. SAGE.

Cunningham, W. (1992). A WyCash portfóliókezelő rendszer. In *Addendum to the Proceedings on Object-Oriented Programming Systems, Languages, and Applications*, 29-30.

Davenport, T. (2016). A stratégiai gépek felemelkedése. *MIT Sloan Management Review*, *58*(1), 29-30.

Desai, D. R., & Kroll, J. A. (2017). Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law & Technology, 31*(1), 1-63.

Doshi-Velez, F., & Kim, B. (2017). Az értelmezhető gépi tanulás szigorú tudománya felé. https://arxiv.org/pdf/1702.08608v2.pdf.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Megmagyarázható mesterséges intelligencia: A survey. *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*.

Dourish, P. (2016). Algoritmusok és társaik: Algoritmikus kultúra kontextusban. *Big Data & Society, 3*(2), 1-11.

Edwards, L., & Veale, M. (2017). Az algoritmus rabszolgája: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law and Technology Review, 16*(1), 18-84.

Edwards, P. N. (2018). Asszimilálódtunk: Néhány alapelv az algoritmikus rendszerekről való gondolkodáshoz. *Az IFIP WG 8.2 konferencia*

jegyzőkönyvei.

Az információs rendszerek és a szervezet kölcsönhatásáról szóló munkakonferencia.

Európai Unió. (2016). Az Európai Parlament és a Tanács (EU) 2016/679 rendelete. Az *Európai Unió Hivatalos Lapja, L 119*(1), 1-88.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Munka és szervezés a tanuló algoritmus korában. *Information and Organization, 28*(1), 62-70.

Firth, N. (2019). Az Apple kártyát vizsgálják, mert azt állítják, hogy alacsonyabb hitelkeretet ad a nőknek. *MIT Technology* Review. https://www.technology review.com/2019/11/11/131983/apple-cardis-being-investigated-over-claims-it-gives-women-lower-credit-limits/

Floridi, L. (2011). A negyedik forradalom gyermekei. *Filozófia és technológia*, *24*(3), 227-232.

Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Adatelemzési kompetencia a vállalati döntéshozatali teljesítmény javítására. *The Journal of Strategic Information Systems, 27*(1), 101-113.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2012). A kvalitatív szigor keresése az induktív kutatásban: Megjegyzések a Gioia-módszertanhoz. *Organizational Research Methods, 16*(1), 15-31.

Gregor, S., & Benbasat, I. (1999). Magyarázatok intelligens rendszerekből: Theoretical foundations and implications for practice. *MIS Quarterly, 23*(4), 497-530.

Jarrahi, M. H. (2018). A mesterséges intelligencia és a munka jövője: Az ember és az AI szimbiózisa a szervezeti döntéshozatalban. *Business Horizons, 61*(4), 577-586.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, a kezemben: Ki a legszebb az országban? A mesterséges intelligencia értelmezéseiről, illusztrációiról és következményeiről. *Business Horizons, 62*(1), 15-25.

Keding, C. (2021). A mesterséges intelligencia és a stratégiai menedzsment kölcsönhatásának megértése: A kutatás négy évtizedének áttekintése. *Management Review Quarterly*, *71*(1), 91-134.

Klein, H., & Myers, M. M. D. (1999). Az információs rendszerekkel kapcsolatos értelmező terepkutatások elvégzésének és értékelésének alapelvei. *MIS Quarterly*, *23*(1), 67-93.

Koutsikouri, D., Lindgren, R., Henfridsson, O., & Rudmark, D. (2018). A kiterjesztése. digital infrastruktúrák: A növekedési taktikák tipológiája. *Journal of the Association for Information Systems, 19*(10), 1001-1019.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digitális iker a gyártásban: Kategorikus irodalmi áttekintés és osztályozás. *IFAC-PapersOnLine*, *51*(11), Elsevier, 1016-1022.

Kruchten, P., Nord, R. L., & Ozkaya, I. (2012). Technikai adósság: A metaforától az elméletig és a gyakorlatig. *IEEE Software, 29*(6), 18-21.

Lindebaum, D., Vesa, M., & Den Hond, F. (2020). A "gép megáll" meglátásai az algoritmikus döntéshozatal racionális feltételezéseinek jobb megértéséhez és annak a szervezetekre gyakorolt hatásai. *Academy of Management Review, 45*(1), 247-263.

Linden, A., Reynolds, M., & Alaybeyi, S. (2019). *5 mítosz a megmagyarázható mesterséges intelligenciáról.* Gartner Research.

Lipton, Z. C. (2018). A modell értelmezhetőségének mítosza. *ACM Queue, 16*(3), 1-27.

Liu, N., Du, M., & Hu, X. (2020). Ellenséges gépi tanulás: Egy értelmezési perspektíva. https://arxiv.org/pdf/2004.11488.pdf.

London, A. J. (2019). Mesterséges intelligencia és fekete dobozos orvosi döntések: Pontosság kontra megmagyarázhatóság. *Hastings Center Report, 49*(1), 15-21.

Lyytinen, K., Nickerson, J. V, & King, J. L. (sajtóban). Metahumán rendszerek = ember + tanuló gépek. *Journal of Information Technology*. https://doi.org/10.1177/0268396220915917.

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Az osztályozási modellek teljesítménye a felhasználó szemszögéből. *Decision Support Systems, 51*(4), 782-793.

Martin, K. (2019). Etikus algoritmusok tervezése. *MIS Quarterly Executive, 18*(2), 129-142.

McBride, N., & Hoffman, R. R. (2016). Az etikai szakadék áthidalása: Az emberi elvektől a robotok utasításaiig. *IEEE Intelligent Systems, 31*(5), 76-82.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A. és mások. (2020). Az emlőrákszűrés mesterséges intelligencia rendszerének nemzetközi értékelése. *Nature, 577*(7788), 89-94.

Miles, M. B., Huberman, M. A., & Saldana, J. (2014). Következtetések levonása és ellenőrzése. In *Kvalitatív adatelemzés: A methods sourcebook* (pp. 275-322). SAGE.

Miller, T. (2019). Magyarázat a mesterséges intelligenciában: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Magyarázatok magyarázata a mesterséges intelligenciában. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Mumford, E. (2006). A szociotechnikai tervezés története: Gondolatok a sikerekről, kudarcokról és lehetőségekről. *Information Systems Journal, 16*(4), 317-342.

Myers, M. D., & Newman, M. (2007). A kvalitatív interjú az IS-kutatásban: Examining the craft. *Information and Organization*, *17*(1), 2-26.

Newell, S., & Marabelli, M. (2015). Az algoritmikus döntéshozatal stratégiai lehetőségei (és kihívásai): Felhívás az "adatifikáció" hosszú távú társadalmi hatásaira. *Journal of Strategic Information Systems, 24*(1), 3-14.

O'Neil, C. (2016). *A matematikai pusztítás fegyverei: Hogyan növelik a nagy adatok az egyenlőtlenséget és fenyegetik a demokráciát*. Broadway Books.

Pääkkönen, J., Nelimarkka, M., Haapoja, J., & Lampinen, A. (2020). A bürokrácia mint az algoritmikus rendszerek elemzésének és tervezésének lencséje. *Proceedings of the CHI Conference on Human Factors in Computing Systems (Emberi tényezők a számítástechnikai rendszerekben)*.

Pasquale, F. (2015). *A fekete doboz társadalom.* Harvard University Press.

Patton, M. Q. (2001). *Minőségi értékelési és kutatási módszerek* (3. kiadás). SAGE.

Preece, A. (2018). A "miért" kérdése a mesterséges intelligenciában: Az intelligens rendszerek megmagyarázhatósága - nézőpontok és kihívások. *Intelligent Systems in Accounting, Finance and Management, 25*(2), 63-72.

Raisch, S., & Krakowski, S. (sajtóban). Mesterséges intelligencia és menedzsment: The automation-augmentation paradoxon. *Academy of Management Review.* https://journals.aom.org/doi/10.5465/2018.0072.

Ribera, M., & Lapedriza, A. (2019). Tudunk-e jobb magyarázatokat adni? Javaslat a felhasználóközpontú magyarázható mesterséges intelligenciára. In . In *Joint Proceedings of the ACM IUI 2019 Workshops*.

Robbins, S. (2020). A mesterséges intelligencia és a borítékoláshoz vezető út: A tudás mint az első lépés a mesterséges intelligenciával működő gépek felelős szabályozása és használata felé. *AI & Society*, 25, 391-400.

Robbins, S. (2019). Egy félrevezetett elv egy fogással: Magyarázhatóság a mesterséges intelligencia számára. *Minds and Machines, 29*(4), 495-514.

Rolland, K. H., Mathiassen, L., & Rai, A. (2018). Digitális platformok kezelése felhasználói szervezetekben: A digitális lehetőségek és a digitális adósságok kölcsönhatásai. *Information Systems Research, 29*(2), 419-443.

Rosenfeld, A., & Richardson, A. (2019). Magyarázhatóság ember-ügynök rendszerekben. *Autonomous Agents and Multi-Agent Systems*, *33,* 673-705.

Rudin, C. (2019). Hagyjunk fel a fekete dobozos gépi tanulási modellek magyarázatával a nagy téttel bíró döntésekhez, és használjunk helyette értelmezhető modelleket. *Nature Machine Intelligence, 1*(5), 206-215.

El Saddik, A. (2018). Digitális ikrek: A multimédiás technológiák konvergenciája. *IEEE MultiMedia, 25*(2), 87-92.

Salovaara, A., Lyytinen, K., & Penttinen, E. (2019). Nagy megbízhatóság a digitális szervezésben: A tudatlanság, a keretprobléma és a digitális műveletek. *MIS Quarterly*, *43*(2), 555-578.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). Az IS tudományág szociotechnikai kohéziós tengelye: Történelmi öröksége és folyamatos relevanciája. *MIS Quarterly, 43*(3), 695-719.

Sarker, S., Xiao, X., Beaulieu, T., & Lee, A. S. (2018). Tanulás az első generációs kvalitatív megközelítésekből az IS tudományágban: Evolúciós szemlélet és néhány implikáció a szerzők és az értékelők számára (1/2. rész). *Journal of the Association for Information Systems, 19*(8), 752-774.

Sarker, Saonee, & Sarker, Suprateek. (2009). Az agilitás feltárása elosztott információs rendszereket fejlesztő csapatokban: Egy értelmező tanulmány offshoring-kontextusban. *Information Systems Research*, *20*(3), 440-461.

Scheel, P. D. (1993). Robotika az iparban: A safety and health perspective. *Professional Safety, 38*(3), 28-32.

Schneider, S., & Leyer, M. (2019). Én vagy az informatika? A mesterséges intelligencia alkalmazása a személyes stratégiai döntések delegálásában. *Managerial and Decision Economics, 40*(3), 223-231.

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Magyarázat az esetalapú érvelésben - perspektívák és célok. *Artificial Intelligence Review, 24*, 109-143.

Sousa, W. G. de, Melo, E. R. P. de, Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). Hogyan és hová tart a mesterséges intelligencia a közszférában? szakirodalmi áttekintés és

kutatási menetrend. *Government Information Quarterly, 36*(4), 101392.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Julia, H., Kalayanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *Mesterséges intelligencia és az élet 2030-ban: EgySzáz évestanulmány a mesterséges intelligenciáról: A 2015-2016-os tanulmányozó testület jelentése.* Stanford University. https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0831fnl.pdf

Tavory, I., & Timmermans, S. (2014). *Abduktív elemzés: A kvalitatív kutatás elméletalkotása*. University of Chicago Press.

Weller, A. (2019). Átláthatóság: Motivációk és kihívások. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*.

Wright, S. A., & Schultz, A. E. (2018). A mesterséges intelligencia és az üzleti automatizálás növekvő áradata: etikai keretrendszer kidolgozása. *Business Horizons, 61*(6), 823-832.

A. függelék: A DBA ML-projektjei

A projekt neve	A projekt leírása (felhasználási eset a DBA-n belül és a végfelhasználóknál)	Cél	Bemenet	Kimenet	Modell és eszköz
Könyvvizsgálói nyilatkozat	A könyvvizsgálói nyilatkozat modellje felgyorsítja annak ellenőrzését, hogy a könyvvizsgálói nyilatkozatban szereplő vállalati eszközök értékelése helyes-e, és hogy a nyilatkozat nem tartalmaz-e jogsértéseket. Az algoritmust a DBA belső ügyintézői használják.	A vállalati vagyon téves kimutatásának megelőzése	Az eszközérték eléseket bemutató könyvvizsgálói nyilatkozatok szövegei	A jogsértések valószínűsége az eszközértékelésekben	Véletlen erdő, szavak zsákja
Csőd	A csődmodell a vállalati nehézségeket és a fizetésképtelenséget jelzi előre, és kapcsolódik a korai figyelmeztető európai kezdeményezéshez (EWE). Az algoritmust nem a DBA-nál, hanem az EWE közösség külső tanácsadói használják Dániában és máshol az Európai Unióban. A DBA nem felelős az eszközzel kapcsolatos intézkedésekért és következményekért.	a bajba jutott vállalatok azonosítása, hogy időben be tudjanak avatkozni	A cégnyilvántartásból és az éves beszámolókból származó adatok	A csőd valószínűsége	Scikit-learn, gradiens boosting
Cégbejegyzés	A cégbejegyzési modell célja az újonnan bejegyzett dániai cégek csalásra utaló magatartásának felderítése. Az algoritmust a DBA belső ügyintézői használják.	A társasági formával való visszaélés megakadályozása csalás elkövetése céljából	A cégnyilvántartásból, az éves beszámolókból és a HÉA-jelentésekből származó adatok.	A csalárd cselekmények valószínűsége	XGBoost
Földterület és épületek	A Föld és épületek modell az ingatlanvagyonnal és a hosszú távú befektetésekkel kapcsolatos számviteli politikák megsértését jelzi előre. Az algoritmust a DBA belső területi szakértői használják.	A számviteli politika megsértésének megelőzése	A számviteli politikára vonatkozó szöveg a könyvvizsgálói nyilatkozatból	A számviteli politikák megsértésének valószínűsége	Véletlen erdő, szavak zsákja
Személyazonosság ellenőrzése	A személyazonosító okmányok ellenőrzésének modellje felgyorsítja a benyújtott dokumentumok feldolgozását azáltal, hogy a személyazonosító okmány gépileg olvasható részéből egy szöveges karakterláncot szolgáltat, és azt összehasonlítja a felhasználó által megadott adatokkal. Az algoritmust a DBA belső ügyintézői használják.	A dokumentumok feldolgozásának megkönnyítése	A DBA-hoz benyújtott igazolványok képei	JSON karakterlánc az azonosító gépileg olvasható részének szövegével	PassportEye
Ajánlás	Az Ajánlási modell a személyre szabott tartalomra és az optimalizált felületekre összpontosítva javítja a DBA virk.dk online portáljának felhasználói élményét. Az algoritmus javítja a portál használhatóságát a külső ügyfelek (végfelhasználók) számára.	Az online portál használhatóságának javítása	Telemetriai adatok virk.dk-tól	Releváns tartalom ajánlása	[A tanulmány készítésének időpontjáig nem született döntés]
Szektor kód	Az ágazati kódmodell felgyorsítja a vállalat ágazati kódjának ellenőrzését. Jelenleg a vállalati kódok 25%-a hibás. Az algoritmust a DBA belső ügyintézői használják.	Az iparági és ágazati kódok téves jelentésének megelőzése	Tevékenység-leírás szövege a vállalat éves beszámolójából	Valószínűségi eloszlás az ágazati kódok halmazán	Neurális hálózat

Aláírás	Az aláírásmodell a kapcsolódó dokumentumszűrővel együtt felgyorsítja annak ellenőrzését, hogy egy vállalat által létrehozott dokumentum alá van-e írva vagy sem. A DBA belső ügyintézői által használt algoritmus három valószínűséget ad vissza: azt, hogy a dokumentum fizikailag alá van-e írva, hogy digitálisan van-e aláírva, és hogy hiányzik-e az aláírás.	A vállalatalapítás folyamatának megkönnyítése	Egy cégalapítási dokumentum képe	Annak valószínűsége, hogy egy dokumentumot aláírtak-e vagy sem	Neurális hálózat (ResNet16)

B. függelék: Az interjú protokollja

Személyes háttér

Beszélne nekünk az egyetemi és szakmai hátteréről?

Mióta tagja a DBA-nak, és mióta tölti be jelenlegi pozícióját? Beszélne nekünk azokról a

projektekről, amelyekben részt vesz a DBA-nál?

ML és AI projektek a DBA-nál

Fel tudná sorolni az ML Lab által jelenleg végzett gépi tanulási és mesterséges intelligencia

projekteket? Le tudná írni azokat az ML/AI projekteket, amelyekben részt vesz?

Milyen típusú algoritmusokat és modelleket használnak ezekben a

projektekben? Mi az oka e modellek használatának?

A saját szavaival, meg tudná magyarázni...

- Milyen adatok kerülnek a rendszerbe, és milyen típusú kimenetet biztosít az algoritmus?
- Mennyire érti az algoritmus működését?
- Hogyan értelmezi a kimenetet?

A fekete doboz modellek használata és a megmagyarázhatóság

Mennyire megmagyarázhatóak az Ön által a projektekben használt mesterséges intelligencia döntései?

Ki képes megérteni, hogy a mesterséges intelligencia hogyan állítja elő a kimeneteit (adattudósok, fejlesztők, ügyintézők, ...)?

Találkozott már olyan esettel, amikor meg kellett magyaráznia egy bizonyos mesterséges intelligencia döntést? Le tudná írni az esetet részletesen?

Dokumentálták ezt a magyarázatot? Tudna dokumentumokat benyújtani?

Tudna konkrét példát mondani egy tipikus döntésre, amelyet az Ön

mesterséges intelligenciája hoz? Hogyan magyarázná meg a kapott döntést,

ha erre kérnék...

- Képzett könyvvizsgálók által?
- Egy érintett szervezet által?
- A nagyközönség által?

Mi lenne a magyarázat kérésének és átadásának eljárása?

A magyarázat az algoritmus (vagy az előre meghatározott protokoll) tervezésébe van beágyazva, vagy *ad hoc* / kialakulóban van?

Magyarázhatósági követelmények

Hogyan jelenik meg a megmagyarázhatóság követelménye az algoritmusfejlesztésben?

- Használ különböző gépi tanulási platformokat olyan projektekhez, amelyek megmagyarázható mesterséges intelligenciát igényelnek?

Voltak-e problémái vagy problémái a magyarázhatósággal kapcsolatban (a fejlesztés során, a külső érdekeltekkel való kapcsolatokban, a DBA-n belül, vagy a vezetőkkel kapcsolatban)?

- Kértek magyarázatot? Ki kérte?
- Kérésre kielégítő magyarázatot tudott-e adni?
- Tapasztalta, hogy nem tudott magyarázatot adni egy érdekelt félnek, vagy nem tudott magyarázatot kapni

tőle? Hogyan kell figyelembe venni a magyarázhatóságot a rendszerfejlesztés során?

Milyen tervezési elveket alkalmaztak a PROJECTX fejlesztése során (költség, idő stb.)?

Hogyan szervezték meg a PROJECTX tervezését (vízeséses modell szerint, sprintekben stb.)? A

magyarázhatóság rendszerkövetelmény volt-e a mesterséges intelligencia tervezése során?

- Mit jelentett ez a tervezési folyamatra nézve?

- Ha a megmagyarázhatóságot eredetileg rendszerkövetelményként határozták meg, a végső tervezés során a szándéknak megfelelően valósult-e meg? Azaz, a végső terv magyarázhatósága megfelelt az elképzeléseknek?

Írja le a magyarázat elkészítésének folyamatát:

- Ki hozza létre?
- Milyen gyakran és kinek?
- Mik a lépések?

A tervezési fázisban a tervezési elvek közül volt-e konfliktus a megmagyarázhatósággal?

- Ha igen, hogyan navigáltál a kérdésben?

Észleltek-e konfliktusokat az algoritmus által végzett munka eltérő értelmezésével kapcsolatban?

- Tudna példákat mondani?
- Elfogadható-e ez a konfliktus, vagy az ellentmondásokat össze kell egyeztetni?
- Hogyan egyeztethetők össze?
- Ön szerint mi a legjobb módja a konfliktusok

megoldásának? A **megmagyarázható mesterséges**

intelligencia kifejlesztésének okai és következményei Melyek

a fő okai a mesterséges intelligencia megmagyarázásának?

Miért van szükség a megmagyarázhatóságra?

- Belső célokra: hogy kiderítse, hogyan fejlesztheti a mesterséges intelligenciáját, vagy hogy kétszeresen ellenőrizze a kimeneteit?
- Külső célokra: hogy kormányzati hatóságként elszámoltatható legyen, védhető, elfogulatlan

folyamatokkal? Külső nyomás a magyarázhatóság érdekében:

- Meg kell tudnia magyarázni a mesterséges intelligenciával kapcsolatos döntéseket az ügyfeleknek (adófizetőknek)? Hogyan és milyen részletességgel?
- Milyen előírások, belső irányelvek, külső nyomás stb. kényszerítik Önt arra, hogy megmagyarázza a mesterséges intelligencia döntéseit?
- Kik a főszereplők, akiknek magyarázatokat készít? Meg tudná nevezni őket, és példákat tudna mondani arra, hogy milyenek ezek a magyarázatok?

Hogyan korlátozzák a megmagyarázhatósági követelmények a mesterséges intelligencia fejlesztési folyamatát? Le tudná írni ezeket a korlátozásokat?

- Korlátozza a mesterséges intelligenciával kapcsolatos megközelítések használatát a

megmagyarázhatóság szükségessége miatt? Hogyan befolyásolja a rendszerek teljesítményét

az, hogy magyarázható rendszereket kell létrehozni?

Összességében, hogyan befolyásolja a magyarázhatóság a szervezeti célok elérésének képességét?

C. függelék: A kódolás

Fogalmak (elsőrendű)	Témák (másodrendű)	Összesített méretek	Példa idézetek
• A küszöbértékek ellenőrzése az esetkezelők által • Útmutatás a küszöbértékek meghatározásához • A küszöbértékek függése a kódtól	Küszöbértékek	Határmenti borítékok	"De mi többé-kevésbé végig részt veszünk a folyamatban, mert ha hirtelen probléma merül fel, vagy hirtelen felmerül a 'Oké, ezt be tudjuk vetni, de azt akarod, hogy a gép ezt vagy ezt csinálja? Azt akarod, hogy legyen egy jelölés, amely azt mondja, hogy ez az ügy nem mehet tovább, vagy csak azt akarod, hogy menjen át, és [nekünk] legyen egy speciális jelölés, ahol később utánanézhetünk?".... Tehát végig részt veszünk, de bizonyos pontokon inkább [a célokban vagy a gyakorlatban] segítünk, vagy [megkérdezzük], hogy 'Meg tudjuk-e csinálni...?'".
• A valószínűségek átalakítása zászlókká • A mesterséges intelligencia csak az alapvető hibákat jelzi a dokumentumokban	Zászlók		
• Könnyebben átadható mesterséges intelligencia tervezése • Alapvető AI-eszközök széleskörű alkalmazhatósággal	Egy feladat kisebb részekre való felosztása		
• Az egyszerű algoritmusok könnyű magyarázat • A megmagyarázhatóság és a teljesítmény közötti kompromisszum nem mindig létezik - az egyszerű modellek jól működnek.	Értelmezhető algoritmusok kiválasztása		
• Szoros kommunikációs kapcsolatok a fejlesztés során felmerülő félreértések csökkentése érdekében • Kommunikáció a fejlesztőkkel	Társadalmi párbeszéd		
• A bemeneti adatok megértése fontos • Az inputok minősége	Bemeneti vezérlés	Bemeneti és kimeneti burkológörbék	"Egy példa lehetne, hogy a modellünk [arra vonatkozóan, hogy egy dokumentum alá van-e írva vagy sem], ahogy most is, ha a modell azt jósolja, hogy a dokumentum alá van írva, akkor egy speciális kódot kap: "dokumentum aláírt, minden rendben", és ha nincs aláírva, akkor egy másik jelölést kap: "dokumentum nem aláírt". Ezeket az eseteket végigmegyünk, és akkor láthatjuk, hogy ez helyes volt, és ez nem volt helyes. Ebben az esetben nincs igazán - nem kell tudnunk - nem kell tudnom [az ügyekkel foglalkozó munkatársként], hogy a modell miért mondta azt, hogy "aláírt" vagy "nem aláírt", mert azonnal látom, hogy helyes vagy nem helyes.".
• A megmagyarázhatóság által kiváltott alacsonyabb teljesítmény kompenzálása a kimenet felhasználásának ellenőrzésén keresztül • A fekete doboz elfogadhatósága, ha a kimenetek ellenőrzése egyszerű.	Kimeneti vezérlés		
• Az ellenőrzés mint a bizalom megalapozásának segítője az ML-rendszerben • a végső felelősséget viselő ember • Egyszerű algoritmusok, amelyeket egy emberi szakértő követni és reprodukálni tud.	Emberi ellenőrzés		
• Külső érdekelt felek bevonása a fejlesztés korai szakaszába • Visszajelzési csatornák létrehozása a műszaki és üzleti csapatok között	Emberi visszajelzés	Modellválasztási borítékok	"Körülbelül 160 szabályunk van. Vannak technikai szabályaink, amelyek azt vizsgálják, hogy a megfelelő taxonómiát használják-e, hogy az XBRL formátumú-e, és hogy megfelel-e a követelményeknek. Vannak üzleti szabályaink is. Például, hogy az eszközök és a kötelezettségek egyeznek-e? Néhány szabály csak a technikai kérdéseket vizsgálja a példányjelentésben. Vannak olyan szabályok, amelyeket teljes körű szabályoknak nevezünk: ... a benyújtók nem nyújthatják be a jelentést, amíg ki nem javították a hibát. Vannak több iránymutató jellegű szabályaink is, amelyeknél azt mondjuk: "Úgy tűnik, hogy hibát készül elkövetni. A legtöbb ember így csinálja. Biztos, hogy ezt akarja? folytassa a jelentés benyújtását? És akkor [a
• A mesterséges intelligencia fejlesztésének irányítása • Házon belüli fejlesztés, a jobb megértés érdekében	Folyamatos fejlesztési eljárás		

			felhasználók] dönthetnek úgy, hogy figyelmen kívül hagyják a szabályt."
• A képzési adatok belső felhalmozása • Adatok "vörös heringek" • Házon belüli adatokon alapuló képzés	Az adatok ismerete	Képzési adatok borítékok	"Azt hiszem, fontos, hogy gyakran megnézzük ezeket a modelleket, hogy lássuk, változik-e valami. És esetleg újra kiképezni őket. Mert szerintem lehetnek problémák a robusztussággal. Még nem vezettük be ezt a rendszert a gyártásba, de szerintem már úton van afelé."
• A modellek létrehozásának kihívásai • A nyílt interneten történő modellkiképzés veszélyei • A modellek szakaszos képzése	A modell szakaszos képzése		

A szerzőkről

Aleksandre Asatiani a Göteborgi Egyetem Alkalmazott Információs Technológia Tanszékének információs rendszerekkel foglalkozó adjunktusa. Emellett a Svéd Digitális Innovációs Központ (SCDI) munkatársa. Kutatásainak középpontjában a mesterséges intelligencia, a robotizált folyamatautomatizálás, a virtuális szervezetek és az IS sourcing áll. Munkái korábban olyan vezető IS-folyóiratokban jelentek meg, mint az *Information Systems Journal*, a *Journal of Information Technology* és a *MIS Quarterly Executive*.

Pekka Malo az Aalto University School of Business statisztika professzora. Kutatásait az operációkutatás, az informatika és a mesterséges intelligencia vezető folyóirataiban publikálta. Pekka az egyik úttörőnek számít az evolúciós optimalizációs algoritmusok fejlesztésében a kihívást jelentő kétszintű programozási problémák megoldására. Kutatási területe az üzleti analitika, a számítógépes statisztika, a gépi tanulás, az optimalizálás és az evolúciós számítás, valamint ezek marketingre, pénzügyekre és egészségügyre való alkalmazása.

Per Rådberg Nagbøl a Koppenhágai Informatikai Egyetem doktorandusza, aki a Dán Gazdasági Hatósággal együttműködve doktorál az információs rendszerek területén. Akciótervezési kutatást alkalmaz a gépi tanulás minőségbiztosítására és értékelésére szolgáló rendszerek és eljárások megtervezésére, a pontos, átlátható és felelős használatra összpontosítva a közszférában a kockázatkezelés szempontjából.

Esko Penttinen a helsinki Aalto University School of Business információs rendszerek gyakorlati professzora. Információs rendszerekkel kapcsolatos tudományokból doktorált és a Helsinki School of Economics közgazdaságtanból szerzett MSc diplomát. Esko vezeti a Real-Time Economy Competence Center-t, valamint az XBRL Finland társalapítója és elnöke. Tanulmányozza az ember és a gépek közötti kölcsönhatást, a mesterséges intelligencia szervezeti megvalósítását, valamint a kiszervezéssel és a virtuális szervezéssel kapcsolatos irányítási kérdéseket. Fő gyakorlati szakértelmét a szervezetközi információs rendszerek asszimilációja és gazdasági hatásai adják, olyan alkalmazási területekre összpontosítva, mint az elektronikus pénzügyi rendszerek, a kormányzati jelentéstétel és az elektronikus számlázás. Esko kutatásai olyan vezető informatikai szaklapokban jelentek meg, mint a *MIS Quarterly*, *Information Systems Journal*, *Journal of Information Technology*, *International Journal of Electronic Commerce* és *Electronic Markets*.

Tapani Rinta-Kahila az ausztráliai Queenslandi Egyetem UQ Business School és Australian Institute for Business and Economics posztdoktori kutatója. Az Aalto University School of Business informatikai rendszerekkel kapcsolatos tudományokból szerzett doktori fokozatot. Kutatásai az informatika megszűnésével, a mesterséges intelligencia és az automatizálás szervezeti megvalósításával, valamint az informatika sötét oldalával kapcsolatos kérdésekkel foglalkoznak.

Antti Salovaara az Aalto Egyetem formatervezési tanszékének vezető egyetemi tanára és a Helsinki Egyetem informatikai tanszékének adjunktusa. Az ember és az AI együttműködését, az online trollkodást és a felhasználói tanulmányok módszertanát tanulmányozza. Kutatásait mind az ember-számítógép interakcióval, mind az információs rendszerekkel foglalkozó folyóiratokban és konferenciákon publikálta, többek között a *CHI*, a *Human Computer Interaction* és az *International Journal of Human-Computer Studies*, valamint a *MIS Quarterly* és a *European Journal of Information Systems* folyóiratokban.