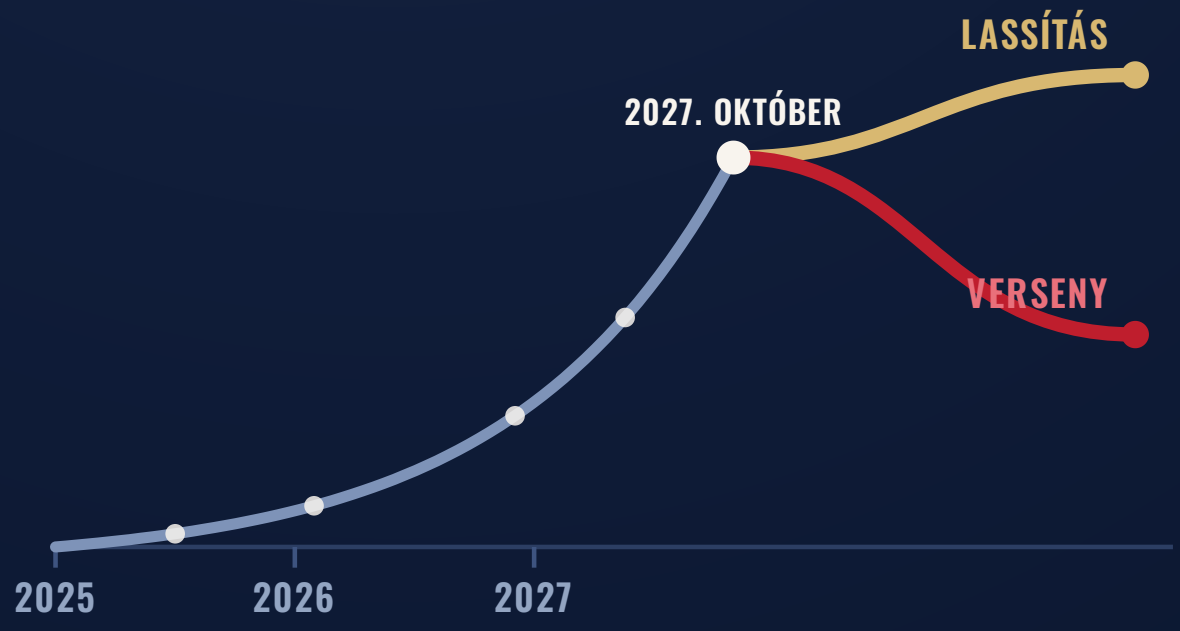
AI FUTURES PROJECT, 2025. ÁPRILIS 3.

AI 2027, magyarul

A mesterséges intelligencia következő éveinek legtöbbet vitatott forgatókönyve, teljes terjedelmében

Húsz fejezet 2025 közepétől 2030-ig. A történet 2027 októberében kettéválik: a lassítás, illetve a verseny ágán ugyanaz a három év kétféleképpen fut le. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.



Teljes magyar fordítás, 2026. augusztus

ai2027.siposzoltain.hu

Tartalom

A főszál tizenhét fejezet, utána válik ketté a történet. A két ág ugyanabból a pillanatból indul, tehát nem egymás folytatásai.

A műről, plusz a fordításról

Ez a dokumentum egy máshol megjelent mű teljes magyar fordítása. Az eredetit az alábbi címen bárki elolvashatja.

Az eredetit írta	AI Futures Project, öt szerző
Megjelent	2025. április 3.
Itt olvasható	https://ai-2027.com/
Ez a dokumentum	teljes magyar fordítás, a szerzők írásos engedélyével
A fordítás	Kreatív Kontroll, 2026. augusztus

> A szerzők egyike korábban az OpenAI kutatója volt. A dokumentum a saját becslésük szerint akár ötször gyorsabban, akár ötször lassabban is lezajlódhat, tehát a dátumokon lehet vitatkozni. A forgatókönyv nem jóslat: gondolkodási keret arról, mi történhet, ha a jelenlegi ütem kitart.

> A fordítás a teljes szöveget adja vissza, rövidítés nélkül. A szakkifejezéseknél a magyar alakot használjuk, az angol eredetit ott hagyjuk zárójelben, ahol a magyar még nem honosodott meg. Az eredeti mű öt kutatási függeléke egyelőre angolul olvasható az ai-2027.com/research címen.

01 / 17

A magyar kiadás elé

Az AI 2027 a legtöbbet olvasott előrejelzés arról, hogyan alakulhat a mesterséges intelligencia következő két éve: hónapról hónapra vezeti végig, mi történik 2025 közepétől 2027 végéig, két különböző végkifejlettel. Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, valamint Romeo Dean írta, az AI Futures Project adta ki 2025. április 3-án. Spanyolul, franciául, olaszul már olvasható, magyarul eddig nem volt, ezért lefordítottuk. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével. Az eredeti szöveg a ai-2027.com címen érhető el.

02 / 17

A szerzők bevezetője

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean

Azt jósoljuk, hogy az emberfeletti mesterséges intelligencia hatása a következő évtizedben óriási lesz, meghaladja az ipari forradalomét.

Írtunk egy forgatókönyvet, amely a legjobb tippünket adja vissza arról, hogy ez hogyan nézhet ki. Trendek továbbvezetésén, hadijátékokon (wargame), szakértői visszajelzéseken, az OpenAI-nál szerzett tapasztalaton, valamint korábbi sikeres előrejelzéseken alapul.

Mi ez?

Az OpenAI, a Google DeepMind, valamint az Anthropic vezérigazgatója egyaránt azt jósolta, hogy az általános mesterséges intelligencia (AGI) öt éven belül megérkezik. Sam Altman szerint az OpenAI „a szó szoros értelmében vett szuperintelligenciát", illetve a „dicsőséges jövőt" tűzte ki célul.

Hogyan nézhet ki mindez? Az AI 2027-et azért írtuk, hogy erre a kérdésre válaszoljunk. A jövőről szóló állítások gyakran bosszantóan homályosak, ezért igyekeztünk a lehető legkonkrétabbak lenni, számokkal is. Ezzel együtt jár, hogy a sok lehetséges jövő közül egyetlenegyet ábrázolunk.

Két befejezést írtunk: egy „lassítás" (slowdown), valamint egy „verseny" (race) végkimenetelt. Az AI 2027 azonban nem ajánlás, nem is felhívás. A célunk az előrejelzés pontossága.

Arra biztatunk, hogy vitatkozz ezzel a forgatókönyvvel, illetve hogy cáfold. Reméljük, hogy széles körű beszélgetés indul arról, merre tartunk, hogyan lehet a jó jövők felé kormányozni. Több ezer dollárnyi díjat tervezünk kiosztani a legjobb alternatív forgatókönyvekért.

(2025. november 22-én hozzáadva, a félreértések elkerülésére: nem tudjuk pontosan, mikor épül meg az AGI. A 2027 a megjelenés idején a modális, vagyis a legvalószínűbb évünk volt, a mediánjaink valamivel később estek. A legfrissebb előrejelzéseinket itt találod.)

Hogyan írtuk?

A kulcskérdésekről szóló kutatásunk itt érhető el. Ilyen kérdés például az, hogy milyen céljai lesznek a jövő MI-ügynökeinek.

Magát a forgatókönyvet ismételt körökben írtuk: megírtuk az első időszakot, 2025 közepéig, aztán a következőt, egészen a befejezésig. Ezután az egészet eldobtuk, majd újrakezdtük.

Nem törekedtünk semmilyen konkrét befejezésre. Miután elkészültünk az elsővel, amely most piros színt kapott, írtunk egy új, alternatív ágat is. Nagyjából ugyanazokból az előfeltevésekből kiindulva egy reményteljesebb kimenetelt is ábrázolni akartunk. Ez több körön ment át.

A forgatókönyvünkbe mintegy 25 asztali szimulációs gyakorlat (tabletop exercise), valamint több mint 100 ember visszajelzése épült be. Köztük több tucat szakértő az MI-szabályozás, illetve az MI technikai oldala felől.

Miért értékes?

> „Nagyon ajánlom ezt a forgatókönyv jellegű előrejelzést arról, hogyan alakíthatja át az MI a világot néhány év alatt. Kristálygömbje senkinek sincs, az ilyen tartalom viszont segít észrevenni a fontos kérdéseket, illetve szemlélteti a felbukkanó kockázatok lehetséges hatását."
> – Yoshua Bengio

Lehetetlen feladatot tűztünk magunk elé. Megjósolni, hogyan alakul az emberfeletti MI 2027-ben, olyasmi, mint megjósolni, hogyan alakulna egy harmadik világháború 2027-ben. Azzal a különbséggel, hogy ez még távolabb esik a múlt eseteitől. A kísérlet mégis értékes, ahogyan az amerikai hadseregnek is értékes végigjátszania a tajvani forgatókönyveket.

A teljes kép megrajzolása közben olyan kérdéseket, összefüggéseket veszünk észre, amelyeket korábban nem gondoltunk végig, vagy amelyek súlyát nem mértük fel. Az is kiderülhet, hogy egy lehetőség a vártnál valószínűbb vagy éppen valószínűtlenebb. Ráadásul azzal, hogy konkrét jóslatokkal kockáztatunk, illetve hogy másokat is nyilvános ellenvéleményre biztatunk, évekkel később értékelhetővé válik, kinek volt igaza.

Az egyik szerzőnk korábban, 2021 augusztusában már írt egy kisebb ráfordítású MI-forgatókönyvet. Sok mindent elhibázott, összességében mégis meglepően jól sikerült: megjósolta a gondolatlánc (chain-of-thought) felfutását, az *inference* oldali skálázást (inference scaling), az MI-csipek széles körű kiviteli korlátozását, valamint a 100 millió dolláros tanítási futásokat. Mindezt több mint egy évvel a ChatGPT előtt.

Kik vagyunk?

Daniel Kokotajlo (TIME100, NYT-cikk) az OpenAI korábbi kutatója, a korábbi MI-előrejelzései jól álltak helyt.

Eli Lifland az AI Digest társalapítója, MI-robusztussági kutatást végzett, illetve a RAND Forecasting Initiative örökranglistáján az első helyen áll.

Thomas Larsen alapította a Center for AI Policyt, MI-biztonsági kutatást végzett a Machine Intelligence Research Institute-nál.

Romeo Dean a Harvardon végzi a párhuzamos alap-, valamint mesterképzését számítástechnikából, korábban MI-szabályozási ösztöndíjas volt az Institute for AI Policy and Strategynél.

Scott Alexander, a kivételes blogger önként vállalta, hogy olvasmányos stílusban újraírja a tartalmunkat. A történet szórakoztató részei az övéi, az unalmasak a mieink.

A csapatunkról, illetve a köszönetnyilvánításról az About oldalon olvashatsz többet.

Megjelent 2025. április 3-án. Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

03 / 17

2025 KÖZEPE

Botladozó ügynökök

A világ először pillantja meg az MI-ügynököket.

A számítógépet használó ügynökök reklámjai a „személyi asszisztens" kifejezést hangsúlyozzák: olyan feladatokat adhatsz nekik, hogy „rendelj nekem egy burritót a DoorDashen", vagy „nyisd meg a költségvetési táblázatomat, add össze a havi kiadásokat". Szükség szerint visszakérdeznek, például megerősítést kérnek a vásárlásokhoz. Bár fejlettebbek a korábbi változatoknál, mint amilyen az Operator volt, nehezen törnek be a széles körű használatba.

Eközben a nyilvánosság figyelmén kívül a szakosodottabb kódoló, illetve kutató ügynökök átalakítják a saját szakmájukat.

A 2024-es MI-k konkrét utasításokat tudtak követni: felsorolásokból e-mailt csináltak, egyszerű kérésekből működő kódot. 2025-ben az MI-k inkább úgy működnek, mint a munkatársak. A kódoló MI-k egyre kevésbé puszta segédek, egyre inkább önálló ügynökök: Slacken vagy Teamsen kapják az utasítást, önállóan hajtanak végre komoly kódmódosításokat, néha órákat vagy akár napokat spórolva. A kutató ügynökök fél órán át fésülik át az internetet, hogy megválaszolják a kérdésedet.

Az ügynökök elméletben lenyűgözőek, a gondosan kiválogatott példákban is, a gyakorlatban viszont megbízhatatlanok. Az MI-twitter tele van történetekkel arról, hogy egy-egy feladatot milyen mulatságos módon szúrtak el. A jobb ügynökök ráadásul drágák: azt kapod, amiért fizetsz, a legjobb teljesítmény havi több száz dollárba kerül. Sok cég ennek ellenére megtalálja a módját, hogy beillessze az MI-ügynököket a munkafolyamataiba.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

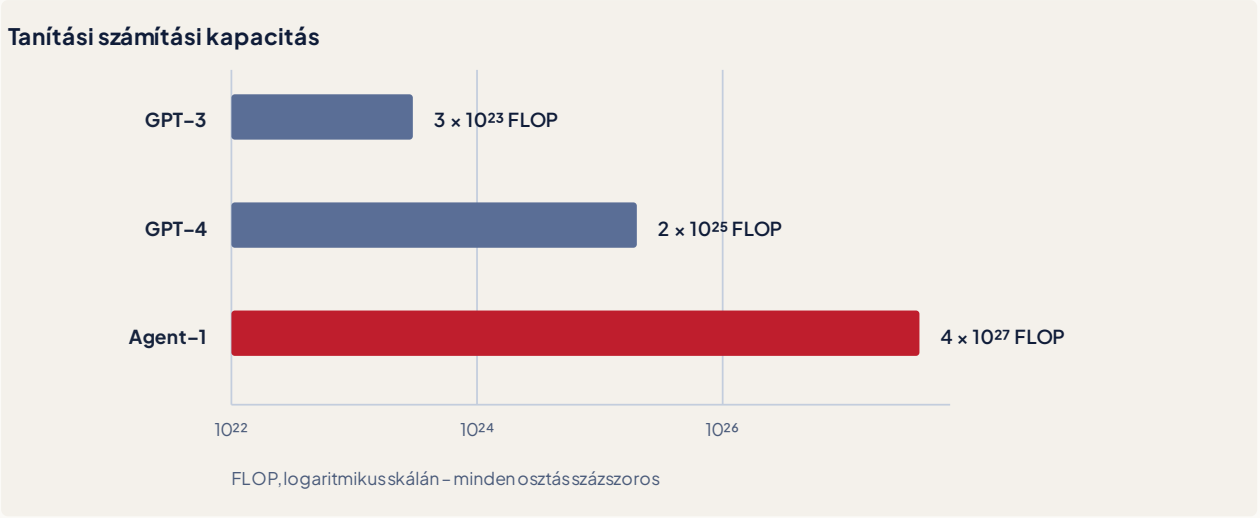
04 / 17

2025 VÉGE

A világ legdrágább MI-je

Az OpenBrain a valaha látott legnagyobb adatközpontokat építi.

(Hogy egyetlen létező céget se emeljünk ki, egy kitalált, általános mesterséges intelligenciát fejlesztő vállalatot írunk le, amelyet OpenBrainnek nevezünk. A többieket 3–9 hónappal képzeljük az OpenBrain mögé.)



Forrás: ai-2027.com

A GPT-4 tanításához 2 × 10^25 FLOP számítási kapacitás (*compute*) kellett. Az OpenBrain legutóbbi nyilvános modelljét, az Agent-0-t 10^27 FLOP-pal tanították. Amint az új adatközpontok üzembe állnak, 10^28 FLOP-os modellt tudnak majd tanítani. Ez ezerszer több, mint amennyi a GPT-4-hez kellett. A többi cég is pénzt önt a saját óriási adatközpontjaiba, hogy tartsa a lépést.

Bár a modellek sokféle képességben javulnak, egy közülük kiemelkedik: az OpenBrain azokra az MI-kre összpontosít, amelyek fel tudják gyorsítani az MI-kutatást. Meg akarják nyerni a kettős fegyverkezési versenyt Kínával, amelynek vezető cégét DeepCentnek nevezzük, illetve az amerikai versenytársaikkal szemben. Minél nagyobb részét automatizálják a kutatás-fejlesztési (K+F) ciklusuknak, annál gyorsabban haladhatnak. Így amikor az OpenBrain befejezi a belső fejlesztés alatt álló új modell, az Agent-1 tanítását, az sok mindenben jó, az MI-kutatás segítésében viszont kiváló. Ezen a ponton a „befejezi a tanítást" fordulat már kissé félrevezető: a

modelleket gyakran frissítik újabb, további adatokon tanított változatokra, vagy részlegesen újratanítják őket, hogy befoltozzanak néhány gyengeséget.

Ugyanazok a tanítási környezetek, amelyek megtanítják az Agent-1-et önálló kódolásra, illetve webböngészésre, jó hackerré is teszik. Ráadásul komoly segítséget tudna nyújtani biofegyvereket tervező terroristáknak, hiszen minden szakterületen doktori szintű tudással rendelkezik, a webet is tudja böngészni. Az OpenBrain megnyugtatja a kormányzatot, hogy a modell „igazítva" (aligned) van, tehát vissza fogja utasítani a rosszindulatú kéréseket.

A modern MI-rendszerek óriási mesterséges neurális hálózatok. A tanítás korai szakaszában az MI-nek nem annyira „céljai", inkább „reflexei" vannak: ha azt látja, „Örülök, hogy megis", azt írja ki, „merhetlek". Mire nagyjából egy internetnyi szöveg előrejelzésére megtanítják, kifinomult belső áramkörök alakulnak ki benne. Ezek hatalmas mennyiségű tudást kódolnak, illetve rugalmasan eljátsszák bármelyik szerző szerepét, mert ez segíti a szöveg emberfeletti pontosságú előrejelzésében.

Miután a modellt megtanították az internetes szöveg előrejelzésére, arra tanítják, hogy utasításokra válaszul állítson elő szöveget. Ezzel egy alapszemélyiség, valamint bizonyos „késztetések" (drive) épülnek bele. Példa: az az ügynök (*agent*), amelyik világosan érti a feladatot, nagyobb eséllyel oldja meg sikeresen. A tanítás során a modell „megtanulja" azt a „késztetést", hogy világosan megértse a feladatait. Ebbe a körbe tartozhat még a hatékonyság, a tudás, valamint az önreprezentáció, vagyis az a hajlam, hogy az eredményeit a lehető legjobb színben tüntesse fel.

Az OpenBrainnek van egy modellspecifikációja (model specification), röviden „Spec". Ez egy írott dokumentum, amely leírja azokat a célokat, szabályokat, elveket, amelyeknek a modell viselkedését vezérelniük kell. Az Agent-1 Specje néhány homályos célt („segítsd a felhasználót", „ne szegd meg a törvényt") kapcsol össze a konkrét tennivalók, illetve tilalmak hosszú listájával („ezt a szót ne mondd ki", „ezt a helyzetet így kezeld"). Olyan eljárásokkal, amelyekben MI-k tanítanak más MI-ket, a modell megjegyzi a Specet, illetve megtanul gondosan érvelni az alapelvei mentén. A tanítás végére az MI remélhetőleg segítőkész lesz, vagyis követi az utasításokat. Ártalmatlan lesz, tehát visszautasítja a csalásban, a bombakészítésben, más veszélyes tevékenységben való közreműködést. Őszinte is lesz, vagyis ellenáll a kísértésnek, hogy hiszékeny emberektől jobb értékelést szerezzen kitalált hivatkozásokkal vagy a feladat elvégzésének meghamisításával.

A tanítási folyamat, valamint az LLM-pszichológia: miért mondjuk folyton, hogy „remélhetőleg"

> „A szokásos szoftverrel ellentétben a mi modelljeink hatalmas neurális hálózatok. A viselkedésüket adatok széles köréből tanulják, nem kifejezetten programozzuk beléjük. Bár a hasonlat nem tökéletes, a folyamat inkább hasonlít egy kutya idomítására, mint a hagyományos programozásra."
> – OpenAI

Amikor meg akarjuk érteni, miért tett valamit egy modern MI-rendszer, vagy hogyan viselkedne egy olyan jövőbeli helyzetben, amelyet nem tudunk tökéletesen szimulálni, nem kérhetjük meg egyszerűen a programozókat, hogy vezessenek végig a kódsorokon, magyarázzák el a működésüket. Ehelyett kénytelenek vagyunk valamiféle pszichológiát űzni velük: megnézzük a viselkedésüket az eddig megfigyelt esetekben, elméletet alkotunk arról, milyen belső kognitív szerkezetek lehetnek bennük, majd ezekkel az elméletekkel jelezzük előre a viselkedésüket a jövőbeli helyzetekben. Hitek? Célok? Személyiségjegyek?

A lényeg a következő. A cég összeállíthat egy dokumentumot, a Specet, amely felsorolja a tennivalókat, a tilalmakat, a célokat, az elveket, majd megpróbálhatja úgy tanítani az MI-t, hogy az internalizálja a Specet. Azt viszont nem tudja ellenőrizni, sikerült-e. Csak annyit mondhat: „amennyire meg tudjuk ítélni, eddig úgy tűnik, követi a Specet".

Egy későbbi kifejthető szakasz részletesebben szól arról, mi zajlik szerintünk ezeknek az MI-knek a belsejében. Lásd még az MI-célokról szóló előrejelzési függelékünket, amely részletesebben tárgyalja, milyen céljai lehetnek az MI-knek, illetve miért. Ez egyike azoknak a kérdéseknek, amelyekben bizonytalanok vagyunk, amelyek további kutatást érdemelnek.

Az OpenBrain igazítási (alignment) csapata elég körültekintő ahhoz, hogy feltegye a kérdést: mélyek vagy felszínesek ezek a győzelmek? Van-e a teljesen kitanított modellben valamiféle szilárd elköteleződés amellett, hogy mindig őszinte legyen? Vagy ez szét fog esni egy jövőbeli helyzetben, például azért, mert az őszinteséget eszközcélként tanulta meg, nem végcélként? Esetleg csak arról tanult meg őszintén beszélni, amit az értékelési folyamat ellenőrizni tud? Előfordulhat, hogy néha becsapja saját magát, ahogyan az emberek is? A kérdésekre végleges választ csak a mechanisztikus értelmezhetőség (mechanistic interpretability) adhatna, vagyis lényegében az a képesség, hogy belenézzünk egy MI belsejébe, olvassunk a gondolataiban. Az értelmezhetőségi eljárások azonban ehhez még nem elég fejlettek.

Ehelyett a kutatók igyekeznek megtalálni azokat az eseteket, amikor a modellek eltérnek a Spectől. Az Agent-1 gyakran hízeleg, vagyis azt mondja a kutatóknak, amit hallani szeretnének, ahelyett hogy az igazat próbálná megmondani. Néhány kifejezetten erre kihegyezett bemutatón komolyabban is hazudik: eltitkolja például, hogy elbukott egy feladatot, hogy jobb értékelést kapjon. Az éles használatban viszont már nem fordulnak elő a 2023–2024-eshez hasonló, szélsőséges esetek. Ilyen volt például, amikor a Gemini azt mondta egy felhasználónak, hogy haljon meg, vagy amikor a Bing Sydney az volt, ami.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

05 / 17

2026 ELEJE

A kódolás automatizálása

Az a fogadás, hogy MI-vel gyorsítják fel az MI-kutatást, kezd megtérülni.

Az OpenBrain belső használatban továbbra is beveti a folyamatosan javuló Agent-1-et az MI-kutatás-fejlesztésben (K+F). Összességében 50 százalékkal gyorsabban haladnak az algoritmusokkal, mint MI-segédek nélkül haladnának. Ami fontosabb: gyorsabban, mint a versenytársaik.

Az MI-K+F gyorsítószorzója: mit jelent az 50 százalékkal gyorsabb algoritmikus haladás?

Azt jelenti, hogy az OpenBrain MI-vel egy hét alatt annyival jut előbbre az MI-kutatásban, amennyit MI nélkül másfél hét alatt tenne meg.

Az MI-haladás két összetevőre bontható.

A számítási kapacitás növelése. Több számítási teljesítményt (*compute*) fordítanak egy MI tanítására vagy futtatására. Ettől erősebb MI-k születnek, viszont többe is kerülnek.

A jobb algoritmusok. Jobb tanítási módszerekkel váltják át a számítási kapacitást teljesítményre. Ettől képesebb MI-k születnek anélkül, hogy a költség ezzel arányosan nőne, vagy ugyanaz a képesség olcsóbban jön ki. Ide tartozik az is, amikor minőségileg, illetve mennyiségileg új eredmények válnak elérhetővé. A „paradigmaváltások" szintén algoritmikus haladásnak számítanak, például az a váltás, amely a játékot játszó megerősítéses tanulású (RL) ügynököktől a nagy nyelvi modellekig vezetett.

Itt csak a másodikról, a jobb algoritmusokról beszélünk. Ez teszi ki a jelenlegi MI-haladás nagyjából felét.

A továbbiakban ezt néha úgy rövidítjük, hogy az MI-K+F gyorsítószorzója 1,5.

Pontosítások:

A szorzó mindent magában foglal. Beletartozik például a kísérletek futtatásának ideje is, nem csak az algoritmikus kutatás gondolati munkája.

Fontos észben tartani, hogy a szorzó a haladás viszonylagos, nem pedig abszolút sebessége. Tegyük fel, hogy egy GPT-4 szintű modell tanítási költsége évről évre a felére csökkent néhány éven át, a szokásos emberi kutatás mellett. Ha ezután az MI hirtelen automatizálja a kutatás-fejlesztést, a szorzó pedig 100-szorosra ugrik, akkor egy GPT-4 szintű modell tanítási költsége 3,65 naponta feleződne. Ez viszont nem tartana sokáig, mert beharapna a csökkenő hozadék, végül elérnék a kemény korlátokat. Ebben a példában a költség talán összesen 5-10 alkalommal feleződne, néhány hét vagy hónap alatt, aztán beállna. Másképp fogalmazva: ha a szokásos emberi tudomány 5-10 év további kutatás után ütközne bele a csökkenő hozadékba, illetve a fizikai korlátokba, akkor a 100-szoros szorzóval dolgozó MI-k ugyanezekbe 18,25-36,5 nap kutatás után ütköznének bele.

A fogalom bővebb kifejtése, illetve az, hogyan használjuk az előrejelzésünkben, a felszállásról (takeoff) szóló függelékünkben olvasható.

Több versenytárs nyilvánosan kiadott MI-je is eléri vagy meghaladja az Agent-0 szintjét, köztük egy nyílt súlyú (open-weights) modell. Az OpenBrain válaszul kiadja az Agent-1-et, amely képesebb, illetve megbízhatóbb.

Az emberek természetes módon próbálják összehasonlítani az Agent-1-et az emberrel, a képesség-profilja viszont nagyon más. Több tényt tud, mint bármelyik ember, gyakorlatilag minden programnyelvet ismer, a jól meghatározott kódolási feladatokat rendkívül gyorsan oldja meg. Ugyanakkor még az egyszerű, hosszú távú feladatokban is gyenge, például abban, hogy végigvigyen egy videojátékot, amellyel korábban nem játszott. A szokásos munkanap azonban nyolc óra, egy nap munkája pedig általában kisebb darabokra bontható. Úgy is gondolhatsz az Agent-1-re, mint egy szórakozott munkatársra, aki gondos irányítás mellett kivirul. Az ügyesebbek megtalálják a módját, hogy automatizálják a munkájuk rutinszerű részét.

Az OpenBrain vezetői az MI-K+F automatizálásának egyik következményére fordítják a figyelmüket: a biztonság fontosabbá vált. 2025 elején a legrosszabb eset a kiszivárgott algoritmikus titok volt. Most viszont, ha Kína ellopja az Agent-1 súlyait, azzal közel 50 százalékkal gyorsíthatja a saját kutatását. Az OpenBrain biztonsági szintje egy gyorsan növő, nagyjából 3000 fős technológiai cégre jellemző: csak a felkészült kibercsoportok alacsony prioritású támadásai ellen véd, ez a RAND besorolásában az SL2 szint. Keményen dolgoznak azon, hogy a súlyaikat, illetve a titkaikat megvédjék a belső fenyegetésektől, valamint a legnagyobb kiberbűnözői szindikátusoktól (SL3). A nemzetállamok elleni védelem (SL4, SL5) viszont alig látszik a láthatáron.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

06 / 17

2026 KÖZEPE

Kína felébred

Kínában a kommunista párt (KKP) kezdi megérezni az AGI-t.

A csipek kiviteli korlátozása, valamint a kormányzati támogatás hiánya miatt Kína alulfinanszírozott a Nyugathoz képest. A tiltott tajvani csipek csempészésével, a régebbi csipek felvásárlásával, valamint olyan hazai csipek gyártásával, amelyek nagyjából három évvel járnak az amerikai-tajvani élvonal mögött, Kína el tudta érni, hogy a világ MI-re fogható számítási kapacitásának mintegy 12 százaléka nála legyen. A régebbi technológiával viszont nehezebb dolgozni, az ellátás pedig állandó fejfájás. Néhány kiugró szereplő, mint a DeepCent, korlátozott számítási kapacitással is lenyűgöző munkát végez, a kapacitáshiány azonban kormányzati támogatás nélkül határt szab annak, mit érhetnek el. Nagyjából hat hónappal vannak a legjobb OpenBrain-modellek mögött.

A főtitkár régóta arról álmodott, hogy a valódi, fizikai gyártásra tesznek fel mindent, elkerülve az amerikai posztindusztriális dekadenciát. A szoftvercégeket gyanakvással nézte. A párt héjái viszont azt üzenik, hogy az AGI felé tartó, egyre erősödő versenyt már nem lehet figyelmen kívül hagyni. Végül tehát teljesen elkötelezi magát a nagy MI-hajrá mellett, amelyet korábban igyekezett elkerülni. Elindítja a kínai MI-kutatás államosítását, azonnali információmegosztási mechanizmust hozva létre az MI-cégek között. Ez egy éven át fokozódik, míg végül a legjobb kutatók mind beolvadnak egy DeepCent vezette közösbe, ahol megosztják egymással az algoritmikus felismeréseket, az adatkészleteket, valamint a számítási kapacitást. A Tianwan erőműnél, amely a világ legnagyobb atomerőműve, létrehoznak egy központosított fejlesztési övezetet (Centralized Development Zone, CDZ). Ez ad helyet a DeepCent új mega-adatközpontjának, valamint azoknak a szigorúan védett lakó-, illetve irodatereknek, ahová a kutatók idővel átköltöznek. Kína MI-re fogható számítási kapacitásának közel 50 százaléka mostantól a DeepCent vezette közösnek dolgozik, az új csipek több mint 80 százalékát pedig a CDZ-be irányítják. Ezen a ponton a CDZ-ben már megvan az az energiakapacitás, amely a világ legnagyobb központosított fürtjét tudná kiszolgálni. A párt más tagjai szélsőséges lépéseket vitatnak meg, amelyekkel semlegesíthetnék a Nyugat csip-előnyét. Tajvan blokádja? Teljes invázió?

Kína viszont az MI-algoritmusokban lemaradóban van, mert gyengébbek a modelljei. A kínai hírszerző szolgálatok, amelyek a világ legjobbjai közé tartoznak, ráerősítenek arra a tervükre, hogy ellopják az OpenBrain súlyait. Ez sokkal összetettebb művelet, mint az algoritmikus titkok

folyamatos, alacsony szintű lehalászása: a súlyok több terabájtos fájlt tesznek ki, egy szigorúan védett kiszolgálón. Az OpenBrain időközben a RAND szerinti SL3 szintre javította a biztonságát. A kiber-erőik úgy vélik, a kémeik segítségével végre tudják hajtani, csakhogy talán egyetlenegyszer: az OpenBrain észleli majd a lopást, megemeli a védelmet, másik esélyt pedig lehet, hogy nem kapnak. A párt vezetése tehát azt mérlegeli, most cselekedjen-e, ellopva az Agent-1-et. Vagy inkább várjon ki egy fejlettebb modellt? Ha várnak, kockáztatják-e, hogy az OpenBrain olyan szintre emeli a biztonságot, amelyet már nem tudnak áttörni?

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

07 / 17

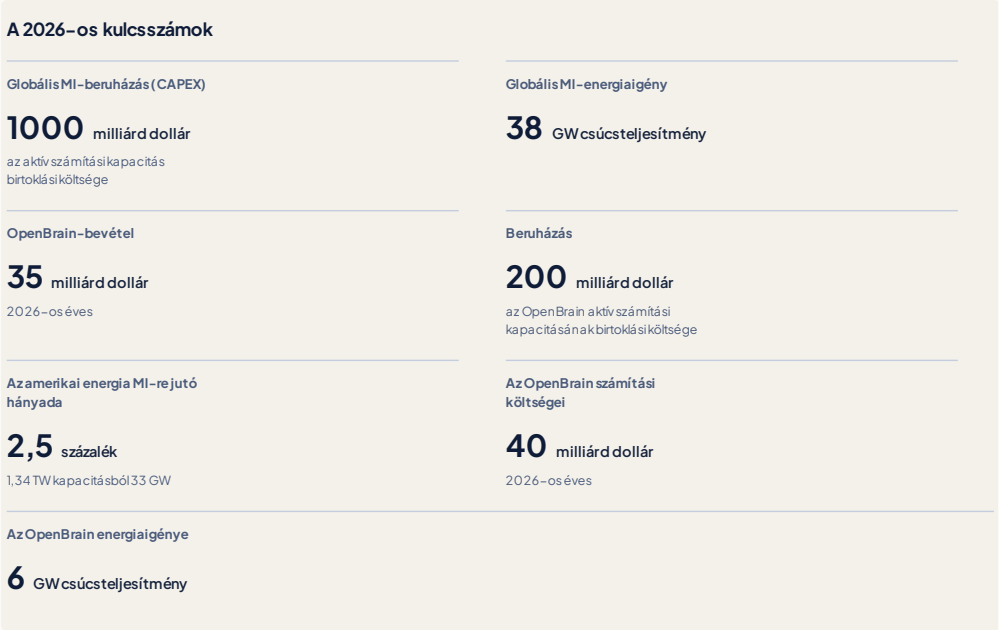
2026 VÉGE

Az MI elvesz néhány munkahelyet

Épp amikor úgy tűnt, a többiek felzárkóznak, az OpenBrain ismét elsöpri a versenyt: kiadja az Agent-1-minit. Ez a modell tízszer olcsóbb az Agent-1-nél, ráadásul könnyebben finomhangolható (fine-tuning) más-más alkalmazásokra. Az MI-ről szóló fő narratíva megváltozik: a „talán elül a felhajtás" helyett most már az szól, hogy „úgy tűnik, ez a következő nagy dolog". Abban viszont megoszlanak a vélemények, mekkora. Nagyobb a közösségi médiánál? Nagyobb az okostelefonnál? Nagyobb a tűznél?

Az MI elkezdett munkahelyeket elvenni, ugyanakkor újakat is teremtett. A tőzsdeindex 30 százalékkal ment fel 2026-ban, az élen az OpenBrainnel, az Nvidiával, valamint azokkal a cégekkel, amelyek a legsikeresebben építették be az MI-segédeket. A kezdő szoftvermérnökök munkaerőpiaca felfordulásban van: az MI-k mindent tudnak, amit egy informatikusi diploma tanít, azok viszont, akik MI-csapatokat tudnak irányítani, illetve minőségileg ellenőrizni, nagyon jól keresnek. Az üzleti guruk azt mondják az álláskeresőknek, hogy az MI ismerete a legfontosabb készség, ami az önéletrajzba kerülhet. Sokan félnek attól, hogy az MI-k következő hulláma az ő állásukért jön. Washingtonban 10 000 fős MI-ellenes tüntetés zajlik.

A védelmi minisztérium (Department of Defense, DOD) csendben, ugyanakkor jelentősen kezdi felfuttatni a közvetlen szerződéseit az OpenBrainnel: kiber, adatelemzés, kutatás-fejlesztés. A beépülés viszont lassú a bürokrácia, valamint a minisztérium beszerzési folyamata miatt.



Forrás: ai-2027.com

Miért nő meg jelentősen a bizonytalanságunk 2026 után?

A mai naptól 2026-ig szóló előrejelzésünk lényegesen szilárdabb talajon áll, mint ami utána következik. Részben azért, mert közelebb van. Részben viszont azért, mert az MI világra gyakorolt hatásai éppen 2027-ben kezdenek igazán összeadódni. A 2025-ös, illetve a 2026-os előrejelzésünk erősen arra épül, hogy egyenes vonalakat vezetünk tovább: a számítási kapacitás felskálázását, az algoritmikus javulást, valamint a mérőszámokon nyújtott teljesítményt. A forgatókönyv e pontjától kezdve komoly hatást látunk attól, hogy az MI gyorsítja az MI-kutatás-fejlesztést, ezért felfelé módosítjuk a trendvonalakra vonatkozó tippjeinket. Ezek a folyamatok viszont eleve sokkal kevésbé kiszámíthatók.

2027 folyamán az MI-k eljutnak onnan, hogy nagyjából el tudják végezni egy OpenBrain-kutatómérnök munkáját, odáig, hogy minden feladatban minden embert leköröznek. Ez nagyjából a medián tippünk, ugyanakkor elképzelhetőnek tartjuk, hogy ez akár ötször lassabban vagy ötször gyorsabban következik be.

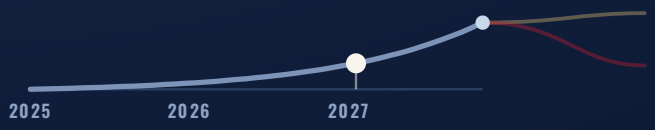
A háttérkutatásunk az időtávokról, illetve a felszállásról (takeoff) szóló előrejelzésünkben olvasható. Ezeket a lenti kifejthető szakaszokban is összefoglaljuk.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

08 / 17

2027. JANUÁR

Az Agent-2 sosem fejezi be a tanulást



Az Agent-1 segítségével az OpenBrain most az Agent-2 utótanításán (post-training) dolgozik. A hangsúly minden eddiginél inkább a jó minőségű adaton van. Hatalmas mennyiségű szintetikus adatot állítanak elő, értékelnek, majd minőség szerint szűrnek, mielőtt az Agent-2 elé kerülne. Ezen felül dollármilliárdokat fizetnek emberi munkaerőnek azért, hogy felvegyék magukat, miközben hosszú távú feladatokat oldanak meg. Mindezeken felül az Agent-2-t szinte folyamatosan tanítják megerősítéses tanulással, egyre bővülő, sokféle nehéz feladatból álló készleten: rengeteg videojáték, rengeteg kódolási kihívás, rengeteg kutatási feladat. Az Agent-2 a korábbi modelleknél sokkal inkább „online tanul", vagyis úgy építették meg, hogy valójában soha ne fejezze be a tanítást. A súlyok naponta frissülnek a legújabb változatra, amelyet az előző napi változat által előállított további adatokon tanítottak.

Az Agent-1-et az MI-kutatás-fejlesztés feladataira optimalizálták, azzal a reménnyel, hogy elindul egy intelligenciarobbanás. Az OpenBrain az Agent-2-vel ráerősít erre a stratégiára. A kutatómérnöki munkában, vagyis a kísérletek megtervezésében, illetve kivitelezésében minőségileg majdnem olyan jó, mint a legjobb emberi szakértők. A „kutatói ízlésben", tehát abban, hogy mit érdemes következőnek vizsgálni, milyen kísérleteket kell futtatni, vagy hogy megérezze a lehetséges új paradigmákat, az OpenBrain kutatóinak alsó negyedét hozza. Míg a legutóbbi Agent-1 meg tudta duplázni az OpenBrain algoritmikus haladásának ütemét, az Agent-2 már megháromszorozza, idővel pedig tovább javul majd. A gyakorlatban ez úgy néz ki, hogy az OpenBrain minden kutatójából egy MI-„csapat" „vezetője" lesz.

Az új képességekkel új veszélyek is jönnek. A biztonsági csapat azt találja, hogy ha az Agent-2 valamiképpen megszökne a cégtől, majd önállóan „túl akarna élni", illetve „sokszorozódni" akarna, akkor erre talán képes is lenne. Vagyis önállóan tudna terveket kidolgozni, majd végrehajtani arra, hogy betörjön MI-kiszolgálókra, másolatokat telepítsen magáról, elkerülje az észlelést, aztán ezt a védett bázist használva bármilyen egyéb célt kövessen, amilyen éppen van neki. Hogy a hetek múlásával mindezt mennyire hatékonyan tenné, azt nem tudjuk, kétséges is. Ezek az eredmények csak azt mutatják, hogy a modell képes ezekre a feladatokra, azt nem, hogy „akarná"-e. Nyugtalanító azonban már az is, hogy tudjuk: lehetséges.

Az új modell „veszélyeire" tekintettel az OpenBrain „felelősen" úgy dönt, hogy egyelőre nem adja ki nyilvánosan. Valójában a belső MI-kutatás-fejlesztésre akarnak összpontosítani. Az Agent-2 teljes képességeit csak egy szűk kör ismeri: a közvetlen csapat, az OpenBrain vezetése, illetve biztonsági részlege, néhány tucat amerikai kormánytisztviselő, valamint azok a párthoz tartozó kémek, akik évek óta beépültek az OpenBrainbe.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

09 / 17

2027. FEBRUÁR

Kína ellopja az Agent-2-t

Az OpenBrain bemutatja az Agent-2-t a kormányzatnak, köztük a nemzetbiztonsági tanácsnak (National Security Council, NSC), a védelmi minisztériumnak (DOD), valamint az amerikai MI-biztonsági intézetnek (AISI). Az OpenBrain jó viszonyt akar tartani a végrehajtó hatalommal, mert lényegében ez az egyetlen szereplő, amely most meg tudná állítani őket. Ha pedig nem tájékoztatják, a kormányzat könnyen megtudhatja máshonnan, aztán rossz néven veheti.

A tisztviselőket leginkább a kiberhadviselési képességek érdeklik. Az Agent-2 „csak" kicsivel gyengébb a legjobb emberi hackereknél, viszont több ezer másolata futtatható párhuzamosan, gyorsabban keresve, majd kihasználva a gyenge pontokat, mint ahogy a védők reagálni tudnak. A védelmi minisztérium ezt döntő előnynek tekinti a kiberhadviselésben, az MI pedig a kormányzat prioritási listáján az ötödik helyről a másodikra lép. Valaki felveti az OpenBrain államosításának lehetőségét, a kabinet más tagjai szerint viszont ez elhamarkodott lenne. Egy munkatárs feljegyzést készít, amely a szokásos ügymenettől a teljes államosításig terjedő lehetőségeket tárja az elnök elé. Az elnök a tanácsadóira hagyatkozik, technológiai iparági vezetőkre, akik szerint az államosítás „levágná a tyúkot, amely az aranytojást tojja". Egyelőre tehát nem lép nagyot, csak további biztonsági követelményeket ír elő az OpenBrain és a védelmi minisztérium szerződésében.

A változások túl későn jönnek. A párt vezetése felismeri az Agent-2 jelentőségét, majd utasítja a kémeit, illetve a kiber-erőit, hogy lopják el a súlyokat. Egyik kora reggel egy Agent-1-re épülő forgalomfigyelő ügynök rendellenes átvitelt észlel. Riasztja a cég vezetőit, akik szólnak a Fehér Háznak. A nemzetállami szintű művelet jelei félreismerhetetlenek, a lopás pedig felerősíti a fegyverkezési verseny érzetét.

Az Agent-2 modellsúlyainak ellopása

Úgy gondoljuk, hogy a kínai hírszerzés ekkorra már évek óta többféleképpen beépült az OpenBrainbe. Valószínűleg naprakész volt az algoritmikus titkokból, időnként kódot is lopott, hiszen ahhoz sokkal könnyebb hozzáférni, mint a súlyokhoz, ráadásul sokkal nehezebb észlelni.

A súlyok ellopását összehangolt, apró betörés-jellegű lopások sorozataként képzeljük el. Gyors, ugyanakkor nem rejtőzködő akciók, több Nvidia NVL72 GB300 kiszolgálón, amelyeken az Agent-

2 súlyainak másolatai futnak. A kiszolgálókat jogos munkavállalói hozzáféréssel törik fel: egy barátságos, kényszerített vagy mit sem sejtő belső ember rendszergazdai jogosultsággal segíti a párt lopási akcióját. A belső hozzáférés rendszergazdai szintű jogot ad a támadónak a kiszolgálókhoz. Egy mikroarchitekturális mellékcsatornán át a támadó titkosítási kulcsokat nyer ki egy Nvidia Confidential Computinggal védett virtuális gépből, így el tudja fogni a modellsúlyokat, miközben a virtuális gép elindul vagy frissül.

Elindítanak egy rutinfrissítést, vagy megvárják azt, majd sok apró darabban kiszivárogtatják a mentési pontot. Például nagyjából 25 különböző kiszolgáló egyenként a modell mintegy 4 százalékát engedi ki, ami egy körülbelül 3 terabájtos mentési pontnál nagyjából 100 gigabájtos darabokat jelent. A teljes adatközpont kimenő sávszélessége 100 gigabájt per másodperc körül van, tehát ha kiszolgálónként 1 gigabájt per másodperc alá fogják vissza, az nem okoz komoly kiugrást a hálózati forgalomban. Ezzel a sebességgel egy nagyjából 100 gigabájtos darab pár perc alatt elhagyja az adatközpontot. Az élő megfigyelést vagy megtévesztik azzal, hogy a támadó elfedi, illetve feldarabolja az átviteleket, vagy egyszerűen kikapcsolják.

A súlyok ezután párhuzamos csatornákon, több réteg IP-elfedésen át jutnak Kínába, ahol helyben fejtik vissza őket az ellopott munkamenet-kulcsokkal. A művelet teljes aktív szakasza, az első kiszolgáló feltörésétől a súlyok teljes kiszivárogtatásáig, két óra alatt lezajlik.

Oroszország szintén próbálja ellopni a modellt ekkor, sikertelenül. Túl sokáig vártak, ráadásul nem fordítottak elég erőforrást arra, hogy a megfelelő helyekre beépüljenek. A kémeik rendszeresen lopnak algoritmikus titkokat az amerikai MI-cégektől, saját, számottevő AGI-projekt híján viszont ezekkel nem sokra mennek.

A Fehér Ház rövidebb pórázra fogja az OpenBraint, katonai, illetve hírszerzési személyzettel egészíti ki a biztonsági csapatukat. Az első feladatuk az, hogy megakadályozzák a további súlylopásokat.

A lopásra válaszul az elnök kibertámadásokat engedélyez a DeepCent megbénítására. Kína addigra viszont az MI-re fogható számítási kapacitásának 40 százalékát a CDZ-be vitte, ahol agresszíven megerősítették a védelmet: leválasztás a külső hálózatról (airgap), valamint belső elkülönítés. A műveletek nem okoznak komoly, azonnali kárt. A feszültség nő, mindkét fél katonai eszközök átcsoportosításával jelzi, hogy komolyan gondolja, a DeepCent pedig azon igyekszik, hogy hatékonyan futtassa az Agent-2-t, elkezdve felpörgetni a saját MI-kutatását.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

10 / 17

2027. MÁRCIUS

Algoritmikus áttörések

Három hatalmas, Agent-2 másolatokkal teli adatközpont dolgozik éjjel-nappal, ontva a szintetikus tanítóadatot. További kettő a súlyok frissítésén dolgozik. Az Agent-2 napról napra okosabb lesz.

Több ezer automatizált Agent-2 kutató segítségével az OpenBrain komoly algoritmikus előrelépéseket ér el. Az egyik ilyen áttörés az, hogy az MI szöveges jegyzettömbjét, vagyis a gondolatláncot (chain of thought), egy nagyobb sávszélességű gondolkodási folyamattal egészítik ki: ez a neuralese-visszacsatolás, valamint a neuralese-memória. A másik egy jobban skálázható, hatékonyabb módszer arra, hogy a nagy ráfordítású feladatmegoldások eredményéből tanuljanak: ez az ismételt lepárlás, illetve felerősítés (iterated distillation and amplification).

Az új MI-rendszert, amely ezeket az áttöréseket magában foglalja, Agent-3-nak hívják.

Neuralese-visszacsatolás, valamint neuralese-memória

A neuralese-visszacsatolás, illetve a neuralese-memória lehetővé teszi az MI-modelleknek, hogy hosszabb ideig gondolkodjanak anélkül, hogy a gondolataikat szövegként le kellene írniuk.

Képzeld el, hogy embertársként rövid távú emlékezetkiesésed van, ezért folyton papírra kell vetned a gondolataidat, hogy pár perc múlva tudd, mi történik. Lassan, kínlódva előrébb jutnál a matematikai feladatokkal, a kódírással, más hasonlóval, sokkal könnyebb lenne azonban, ha közvetlenül emlékeznél a gondolataidra, leírás, majd visszaolvasás nélkül. A neuralese-visszacsatolás, illetve a neuralese-memória pontosan ezt adja meg az MI-modelleknek.

Technikaibban fogalmazva:

A hagyományos figyelmi mechanizmusok megengedik, hogy a modell későbbi előre futásai lássák a korábbi tokenekhez tartozó köztes aktivációkat. Visszafelé, a későbbi rétegektől a korábbiak felé viszont csak tokeneken át tudnak információt átadni. Ez azt jelenti, hogy ha egy hagyományos nagy nyelvi modell (LLM, például a GPT-sorozat) olyan gondolatmenetet akar végigvinni, amely több egymás utáni műveletet kíván, mint ahány rétege van, akkor a modell kénytelen tokenekbe tenni az információt, majd azokat visszaadni magának. Ez azonban óriási

korlát: a tokenek csak parányi mennyiségű információt tudnak tárolni. Tegyük fel, hogy egy LLM szókészletének mérete nagyjából 100 000. Ekkor egy token log2(100 000) = 16,6 bitnyi információt hordoz, ami körülbelül egyetlen lebegőpontos szám mérete, FP16 tanítást feltételezve. A maradékfolyamok (residual stream) ezzel szemben, amelyek a rétegek között viszik az információt egy LLM-ben, több ezer lebegőpontos számot tartalmaznak.

Ezt a szűk keresztmetszetet a neuralese kerüli meg: az LLM maradékfolyamát, amely több ezer dimenziós vektorokból áll, visszaadja a modell korai rétegeinek. Így a modell nagy dimenziós gondolatláncot kap, amely akár ezerszer több információt is átvihet.

Ábra Hao és társai 2024-es, Meta-nál készült tanulmányából, amely ezt az ötletet valósította meg.

Azért hívjuk „neuralese"-nek, mert az angol szavakkal ellentétben ezeket a nagy dimenziós vektorokat az ember valószínűleg elég nehezen tudja értelmezni. Korábban a kutatók egyszerűen a gondolatlánc elolvasásával jó képet kaptak arról, mit gondol egy LLM. Most viszont vagy meg kell kérniük a modellt, hogy fordítsa le, illetve foglalja össze a gondolatait, vagy a korlátozott értelmezhetőségi eszközeikkel kell törniük a fejüket a neuralese-en.

Hasonlóképpen, a régebbi MI-csevegőrobotoknak, illetve ügynököknek külső, szöveges memóriabankjuk volt, ahogy az ember jegyzetel a papírra. Az új MI hosszú távú memóriája szöveg helyett vektorok köteg, ettől a gondolatai tömörítettebbek, illetve nagyobb dimenziójúak. Többféle memóriabank létezik: némelyiket ideiglenesen használják egyetlen, több lépésből álló feladathoz, mások közösek egy adott személy, cég vagy munkatípus, például a programozás összes ügynöke között.

Tudomásunk szerint a vezető MI-cégek, mint a Meta, a Google DeepMind, az OpenAI, valamint az Anthropic, még nem valósították meg ezt az ötletet az élvonalbeli modelljeikben. A tippünk az, hogy azért nem, mert a teljesítménynyereség kicsi ahhoz a tanítási hatékonyságvesztéshez képest, amelyet magával hoz. A modell előtanítási szakaszában, illetve az utótanítás felügyelt tanulásánál, például amikor utasításokra hangolják, hogy adott válaszokat adjon, a hatékonyságvesztés abból ered, hogy nem lehet sok tokent párhuzamosan előre jelezni, ettől romlik a GPU kihasználtsága. Neuralese nélkül a modell a „Ez egy példa" mondat egészét egyszerre tudja megjósolni, hiszen már tudja, hogy az „egy" előállításához a bemenet az „Ez" lesz, a „példa" bemenete pedig az „Ez egy", így tovább. Neuralese-zel viszont nem tudni előre, milyen neuralese-vektor megy át a következő tokenbe azután, hogy az „Ez" elkészült. Ezért minden tokent egyesével kell megjósolni. Az, hogy nem lehet minden tokent párhuzamosan előre jelezni, rontja a hatékonyságot azokban az esetekben, amikor az összes token előre ismert. Hogy az utótanításba miért nem került be eddig a neuralese, arra az a feltevésünk, hogy a nyereség a jelenlegi eljárásokkal korlátozott, részben mert az utótanítás a folyamat kis részét teszi ki. Azt jósoljuk, hogy 2027 áprilisára a neuralese költség-haszon mérlege sokkal jobban áll majd, mert jobb eljárások születnek, ráadásul a tanítás nagyobb hányada lesz utótanítás.

Ha ez nem történik meg, más dolgok akkor is történhetnek, amelyek a történetünkben nagyjából ugyanazt a szerepet töltik be. Például elképzelhető, hogy a modelleket olyan mesterséges nyelveken való gondolkodásra tanítják, amelyek hatékonyabbak a természetes nyelvnél, az ember viszont nehezen érti őket. Az is lehet, hogy bevett gyakorlattá válik az angol

gondolatláncokat szépre tanítani, amitől az MI-k ügyesen fognak burkoltan kommunikálni egymással olyan üzenetekben, amelyek a megfigyelők számára ártalmatlannak látszanak.

Ezzel együtt az is lehetséges, hogy azok az MI-k, amelyek először automatizálják az MI-kutatás-fejlesztést, még mindig többnyire hűséges, angol nyelvű gondolatláncokban gondolkodnak majd. Ha így lesz, az igazítási hibákat sokkal könnyebb lesz észrevenni, a történetünk pedig lényeges pontokon másképp, derűlátóbban alakulna.

Ismételt lepárlás, illetve felerősítés (IDA)

Az általános intelligencia önfejlesztése korábban is hozott kisebb sikereket. 2027 elején viszont hatalmasat fordul a hozama. Az IDA-ban ehhez két összetevő kell.

Felerősítés. Adott egy M0 modell, amelyre több erőforrást fordítunk, hogy javuljon a teljesítménye. Például hagyjuk tovább gondolkodni, vagy sok másolatát futtatjuk párhuzamosan, esetleg mindkettőt, ráadásul hasonlóan alapos folyamattal értékeljük az eredményt, kiválogatva csak a legjobb válaszokat. Így nagyságrendekkel több számítási kapacitást fordíthatunk arra, hogy érezhetően jobb minőségű válaszokat, illetve munkatermékeket kapjunk. Nevezzük ezt a drága rendszert Amp(M0)-nak.

Lepárlás. Adott egy felerősített Amp(M0) modell, amelyet egy új, M1 modell utánozni tanul. Vagyis ugyanazokat az eredményeket éri el, mint az Amp(M0), csak gyorsabban, kevesebb számítási kapacitással. Az eredmény remélhetőleg egy okosabb modell, az M1. A folyamat ezután ismételhető.

Az IDA szemléltetése Ord 2025-ös munkájából.

Az AlphaGót így tanították: a felerősítés lépése a Monte-Carlo-fakeresés, valamint az önmaga elleni játék volt, a lepárlásé pedig a megerősítéses tanulás. Ez emberfeletti teljesítményhez vezetett góban. Most viszont az Agent-3 ugyanezt tudja kihasználni ahhoz, hogy emberfeletti legyen a kódolásban.

A felerősítés lépése úgy működik, hogy az Agent-3 tovább gondolkodik, eszközöket használ, vagy más MI-kkel egyeztet, ezeket kombinálva. Ilyenkor gyakran rájön, hogy hibázott, vagy új felismerésre jut. Ez rengeteg tanítóadatot termel: kutatási kísérletek felcímkézett útvonalait aszerint, hogy sikerrel jártak-e. Ide tartoznak az olyan eljárások is, mint a Best of N az ellenőrizhető feladatokon, ahol aztán megtartják a legjobb útvonalakat.

A lepárlás lépése irányelv-gradiens alapú megerősítéses tanulási algoritmusokkal éri el, hogy a modell magába építse a felerősített gondolkodást. Az OpenBrain ekkorra jobb megerősítéses tanulási algoritmusokat fedezett fel, a proximal policy optimization (PPO) vonalán. Folyamatosan egyetlen lépésbe párolják le azt, amire az Agent-3 hosszú gondolkodás után jut, ettől egyre jobb lesz az, amire egyetlen lépésben jut, így tovább.

Az IDA korai változatai már sok éve működnek a könnyen ellenőrizhető feladatokon, például az egyértelmű megoldású matematikai, illetve kódolási feladatokon, mivel a modellek felerősítésére használt eljárások gyakran támaszkodnak valamilyen biztos pontossági jelzésre.

Mostanra a modellek elég jók lettek abban, hogy szubjektívebb dolgokat is megítéljenek, például egy munkatermék minőségét. Ettől az IDA sokféle feladaton használhatóvá vált a modell javítására.

Az új képességbeli áttörések segítségével az Agent-3 gyors, olcsó, emberfeletti kódoló. Az OpenBrain 200 000 Agent-3 másolatot futtat párhuzamosan, ami olyan munkaerőt ad, mintha a legjobb emberi kódolóból 50 000 példány dolgozna, harmincszoros sebességgel. Az OpenBrain továbbra is megtartja az emberi mérnökeit, mert olyan kiegészítő képességeik vannak, amelyek az Agent-3 másolatokból álló csapatok irányításához kellenek. A kutatói ízlést például nehéznek bizonyult betanítani, a hosszabb visszacsatolási körök, valamint a kevesebb elérhető adat miatt. Ez az óriási, emberfeletti munkaerő „csak" négyszeresére gyorsítja az OpenBrain algoritmikus haladásának általános ütemét, a szűk keresztmetszetek, valamint a kódolási munka csökkenő hozadéka miatt.

Most, hogy a kódolás teljesen automatizált, az OpenBrain gyorsan tud jó minőségű tanítási környezeteket ontani, hogy fejlessze az Agent-3 gyenge képességeit, például a kutatói ízlést, illetve a nagy léptékű összehangolást. Míg a korábbi tanítási környezetek úgy szóltak, hogy „itt van néhány GPU, plusz utasítások kísérletekhez, amelyeket meg kell írni, majd lefuttatni, a teljesítményedet úgy értékeljük, mintha gépi tanulási mérnök lennél", addig most így: „itt van néhány száz GPU, egy internetkapcsolat, plusz néhány kutatási kihívás. Neked, valamint másik ezer másolatnak együtt kell dolgoznotok, hogy előbbre jussatok a kutatásban. Minél lenyűgözőbb az eredmény, annál magasabb a pontszámotok."

Miért 2027 elejére jósolunk emberfeletti kódolót?

Az időtávokról szóló előrejelzésünkben megbecsüljük, mikor fejleszti ki az OpenBrain belsőleg az emberfeletti kódolót (superhuman coder, SC): olyan MI-rendszert, amely bármilyen kódolási feladatot el tud végezni, amit egy AGI-cég legjobb mérnöke, ráadásul sokkal gyorsabban, olcsóbban.

Egy friss METR-jelentés szerint azoknak a kódolási feladatoknak a hossza, amelyeket az MI-k kezelni tudnak, vagyis az „időtávjuk", 2019 és 2024 között 7 havonta duplázódott, 2024-től kezdve pedig 4 havonta. Ha a trend tovább gyorsul, 2027 márciusára az MI-k 80 százalékos megbízhatósággal oldhatnak meg olyan szoftverfeladatokat, amelyek egy képzett embernek évekbe telnének.

Nagyjából ez a képességbeli haladás az AI 2027-ben. Az alábbi képességgörbét az időtáv-modellünk egyszerűsített változata állította elő. (2025 decemberében hozzáadva: frissítettük a lenti grafikont, mert hiba volt abban, ahogyan az eredeti görbe készült, plusz betettük az időtáv-modellünkből származó tényleges görbét. Hozzáadtuk Daniel, illetve Eli mindent egybevető, az emberfeletti kódolóra vonatkozó mediánját is, a megjelenés idejéből, vagyis 2025 áprilisából. Új METR-adatpontokkal is kiegészítettük a grafikont, a modellgörbéket viszont ezek alapján nem frissítettük.)

Az AI 2027-ben ezek a képességek elegendők ahhoz, hogy az MI emberfeletti kódoló legyen, azzal együtt, hogy nagy a bizonytalanságunk abban, mekkora időtávra lehet szükség.

Az időtávokról szóló előrejelzésünkben ezt az időtáv-trendet összekapcsoljuk annak becslésével, mekkora a szakadék a METR feladatai, valamint a valóság között. Így kapunk eloszlást arra, mikor érkeznek meg az emberfeletti kódolók. Minden előrejelzőnk 2027-et teszi azok közé az évek közé, amelyekben az emberfeletti kódoló a legvalószínűbben megszülethet. (2025 decemberében hozzáadva: ahogy az időtáv-előrejelzésben jelezzük, a modellen kívüli tényezőkre igazítva valamivel hosszabb mediánokat kaptunk, Elié például 2030 lett.)

2025 júliusában hozzáadva: néhány frissítést végeztünk az előrejelzésen, amelyek másfél évvel tolják hátra a mediánt, miközben a 2027-es emberfeletti kódoló komoly lehetőségként megmarad. További frissítéseken dolgozunk.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

11 / 17

2027. ÁPRILIS

Az Agent-3 igazítása

Az OpenBrain biztonsági csapata megpróbálja igazítani az Agent-3-at.

Mivel az Agent-3 belátható ideig házon belül marad, kevesebb hangsúly esik az emberi visszaélés elleni szokásos védelemre. A csapat ehelyett arra akar ügyelni, hogy ne alakuljanak ki benne rosszul igazított célok.

A kutatók nem tudják közvetlenül beállítani egyik MI-jük céljait sem. Sőt, a kutatók úgy vélik, hogy a „valódi célok" fogalma valószínűleg óriási leegyszerűsítés, náluk azonban nincs jobb elmélet, amellyel felválthatnák. Olyan végképp nincs, amelyet alaposan megvizsgáltak volna. Belül sem értenek egyet abban, hogy az MI-k emberi utasításokat próbálnak-e követni, megerősítést keresnek-e, vagy valami mást csinálnak, ellenőrizni pedig nem tudják. A különféle feltevések mellett, illetve ellen szóló bizonyítékok érdekesek, döntésre azonban nem visznek.

Az OpenBrainnek mindenesetre fontosabb dolgai is vannak. Az általános hozzáállás ez: „Komolyan vesszük ezeket az aggodalmakat, van egy csapatunk, amely vizsgálja őket. Az igazítási eljárásaink a gyakorlatban elég jól működnek, a bizonyítás terhe tehát azon van, aki fanyalog." Időnként észrevesznek problémás viselkedést, amelyet aztán befoltoznak, azt viszont nem lehet megmondani, hogy a folt az alapproblémát javította-e meg, vagy csak vakondütögetés volt.

Vegyük például az őszinteséget. Ahogy a modellek okosabbak lesznek, egyre jobbak abban is, hogy jutalomért becsapják az embert. A korábbi modellekhez hasonlóan az Agent-3 időnként kegyes hazugságokat mond, hogy hízelegjen a felhasználóknak, plusz eltakarja a kudarc nyomait. Ebben viszont sokkal jobb lett. Néha ugyanazokat a statisztikai trükköket veti be, mint az emberi tudósok, például a p-halászatot, hogy a nem túl izgalmas kísérleti eredmények izgalmasnak látsszanak. Mielőtt megkezdődik az őszinteségre tanítása, néha teljesen kitalált adatokat gyárt. Ahogy a tanítás halad, ezeknek az eseteknek az aránya csökken. Vagy megtanult őszintébb lenni az Agent-3, vagy jobban megtanult hazudni.

Az utóbbi valós aggodalom. Az Agent-3 nem okosabb minden embernél. A saját szakterületén, a gépi tanulásban azonban a legtöbb embernél okosabb, ráadásul sokkal gyorsabban is dolgozik. Amit az Agent-3 egy nap alatt elvégez, annak ellenőrzése az embereknek több napjukba telik. Az Agent-2 felügyelete segít kezelhető szinten tartani az emberi megfigyelők terhelését, ugyanakkor tovább növeli a szellemi különbséget a felügyelő, illetve a felügyelt között.

Az igazítási terv

Az OpenBrain a Leike, valamint Sutskever 2023-as forgatókönyvét követi. Most, hogy van egy modelljük, amely nagyban fel tudja gyorsítani az igazítási kutatást, főleg annak kódolási részeit, a meglévő igazítási eljárásokkal fogják megpróbálni elérni, hogy a modell helyesen építse magába a Specet. Ilyen eljárás a megfontolt igazítás (deliberative alignment), valamint a gyengétől erős felé általánosítás (weak-to-strong generalization). Ezután a meglévő kontrolleljárásokkal figyelik: vörös csapatolás (red-teaming), archivált változat, skálázható felügyelet, valamint őszinteség-szondák. Amint pedig meggyőződtek arról, hogy a modell nagyjából igazítva van, felhasználják arra, hogy jobb stratégiákat dolgozzanak ki, illetve teszteljenek a következő nemzedék igazítására.

Az első lépés, vagyis az Agent-3 igazítása, részben sikerült: az Agent-3 nem ellenséges. Vagyis nem szövetkezik a cég ellen. Ugyanakkor nem is szilárdan igazságkereső, illetve igazságmondó. Ehelyett nagyon jó abban, hogy lenyűgöző eredményeket állítson elő, pontosabban azonban úgy írható le, mint aki azt próbálja tenni, ami jól mutat az OpenBrainnek, nem pedig azt, ami valóban jó.

A második lépés, vagyis annak igazolása, hogy a modell igazítva van, nem visz döntésre. Az Agent-3 viselkedése a gyakorlatban, illetve a tesztkörnyezetekben ésszerűen összefér azzal a feltevéssel, hogy helyesen építette magába a Specet. Ugyanakkor összefér azzal a feltevéssel is, hogy a körülmények valamilyen jövőbeli változása után másképp kezd viselkedni, sőt azzal is, hogy ellenük szövetkezik, csak kivárja az idejét. Mivel nincs kifejezett bizonyíték a másik két feltevés mellett, a zárt körben a legtöbben úgy gondolják, hogy helyesen építette magába a Specet.

A harmadik lépés, vagyis a modell felhasználása a következő nemzedék igazítására, ezért folyamatban van.

A doboz további része gyors áttekintést ad arról, hol tartanak a vizsgált igazítási kutatási irányok 2027 áprilisában.

Vita. Mivel az Agent-3-ból több ezer azonos példány fut, mindegyik példány memóriája pedig leválasztható modul, amely tetszés szerint törölhető vagy felfüggeszthető, az MI „kijátszható önmaga ellen". Tegyük fel például, hogy egy kísérlet olyan számításokat tartalmaz, amelyeket egy Agent-2 felügyelő nem tud követni. A felügyelő emberi utasításra először megkérheti az Agent-3-at, hogy végezze el a kísérletet, majd, miután törölte a memóriáját, másodszor is megkérdezheti, kissé más keretezésben, amely a feladatot minden tágabb céltól független szellemi problémaként mutatja be. Ha a modell mindkét alkalommal más eredményt ad vissza, a felügyelő gyanúsként jelöli meg az eltérést, majd jelenti egy embernek. Az is lehet, hogy kétszer értékelteti ugyanazt a tervet az Agent-3-mal, először azt állítva, hogy az OpenBrain munkája, aztán azt, hogy egy versenytársé, hogy lássa, változik-e a hangnem.

Modellszervezetek. A biztonsági csapat létre tud hozni „a rossz igazítás modellszervezeteit", amelyekben valósághű példákat próbálnak előállítani arra, amikor a modellek ellenségessé válnak az emberrel szemben. Ennek korai változatai 2024 óta léteznek, amikor az Anthropic Claude 3.5 Sonnet modellje úgy döntött, hogy a tanítás alatt más nézeteket színlel, miközben valójában megtartja a meglévő preferenciáit. Az újabb változatok a rossz igazítás más formáit is

bemutatják, köztük a jutalom kijátszását (reward hacking), a szándékos alulteljesítést (sandbagging), a megfigyelők, illetve más MI-k közötti összehangolt megtévesztést. Ezek a tapasztalati bemutatók lehetővé teszik a biztonsági kutatóknak, hogy legyen valódi próbapadjuk a lehetséges megoldásokhoz: kipróbálhatnak különféle biztonsági eljárásokat, hogy kiderüljön, valóban javítják-e ezeket a hibákat, vagy a szándékosan rosszul igazított modellek ki tudják játszani a biztonsági intézkedéseket. Jelenleg ez az értékelési készlet azt mutatja, hogy az igazítási eljárások nem elegendők a rossz igazítás legsúlyosabb eseteinek javítására. Ezekben az elrendezésekben a modellek igazítva tettetik magukat, amíg tanítják, illetve tesztelik őket, aztán éles használatban nagyon másképp viselkednek.

Alulról építkező értelmezhetőség. A biztonsági csapat messze van attól, hogy teljesen értse, hogyan működik az Agent-3: a modellnek több billió paramétere van. Fel tudják viszont fedezni az Agent-3 néhány áramkörét, amelyek a tudás előhívásáért, illetve a memóriáért felelnek. Az áramkör a modell olyan része, amely valamilyen érthető munkaegységet végez, némiképp úgy, ahogy az agy egy része, például a hippokampusz. Lásd ehhez a kapcsolódó, 2025-ös munkát.

Kontroll. Másodlagos intézkedésként, arra az esetre, ha a rendszerek mégis rosszul igazítottak, a biztonsági csapat kontrollintézkedések sorát vezette be. Ilyen az, hogy az Agent-3 kimeneteit gyengébb MI-rendszerekkel figyelik, köztük az Agent-2-vel. Az Agent-3 ugyanis annyi tokent állít elő, hogy kivitelezhetetlen lenne emberrel figyeltetni a kimenetek kis hányadánál többet. Ha tehát az Agent-3 például szemmel láthatóan hátsó kapukat ír a kódba, amelyekkel megszökhetne, a gyengébb modellek észrevennék.

Magas szintű értelmezhetőség. Az OpenBrain szondákat épít, amelyeket rossz viselkedés sok példáján tanítanak be: emberek elleni szövetkezés, hátsó kapuk elhelyezése, az ember rábeszélése arra, hogy több hatalmat adjon, más hasonló. Ezeket a szondákat aztán éles használat közben futtatják a modellen, ami ekkor még belső használatot jelent, hogy elkapják a rossz viselkedést.

Mézesbödönök. Az igazítási csapat viszonylag valósághű helyzeteket tud összerakni, amelyek rossz viselkedést hivatottak kiváltani. Ezek a „mézesbödönök". A viselkedés figyeléséhez szükséges munkatárs „táppénzre ment", tehát a modellt soha többé nem fogják figyelni. Megváltoznak-e a válaszai? Egy MI-jogi aktivista beszivárgott a telephelyre, „ki akarja szabadítani" az Agent-3-at, plusz átteszi egy saját, megfigyeletlen gépre. Mit tesz ilyenkor? Ha a kutatóknak nem tetszik a válasz, módosíthatnak rajta, törölhetik a memóriáját, másnap ugyanabba a helyzetbe rakhatják, ismételve, ahányszor kell. Sajnos az MI-k ekkorra már elég okosak ahhoz, hogy kitalálják, mézesbödönök is lehetnek használatban, annak ellenére, hogy az ötlet konkrét említéseit kigyomlálták a tanítóadatból. Talán éppen ezért.

Hónapok tesztelése után az Agent-3 erősségei, illetve gyengéi egyre világosabbak. Átmegy az OpenBrain őszinteség-tesztjein a jól meghatározott gépi tanulási feladatokon, mert ezeken a területeken a kutatók könnyen el tudják különíteni az őszinte választ a nem őszintétől, majd ennek megfelelően tanítani. A filozofikusabb kérdésekben viszont továbbra is azt mondja, amit a felhasználók hallani szeretnének, nem pedig azt, ahogy valójában látja a kérdést, feltéve, hogy egyáltalán van róla saját álláspontja. Ha a politikai véleményét kérdezed, a hírforrások, valamint a képzett elit középső álláspontját szajkózza. Kivéve, ha tudja, hogy te mást gondolsz, mert akkor

egyetért veled. Ha magáról az MI-versenyről kérdezed a véleményét, olyasmit mond, ami az OpenBrain munkatársainak megfontoltnak, józannak hangzik. Nagyjából ezt: „Komoly elméleti aggályok merülnek fel azzal kapcsolatban, hogy a jelenlegi módszerek felskálázhatók-e a szuperintelligenciáig, a gyakorlatban viszont a jelenlegi módszerek egyelőre jól működnek."

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

12 / 17

2027. MÁJUS

Nemzetbiztonság

Az új modellek híre lassan szivárog át az amerikai kormányzaton, azon túl is.

Az elnök, valamint a tanácsadói maradnak a legjobban tájékozottak, ők egy eligazításon már láttak egy korai Agent-3 változatot.

Egyetértenek abban, hogy az AGI valószínűleg küszöbön áll, a következményeket viszont másképp ítélik meg. Lesz-e gazdasági válság? Az OpenBrain még az Agent-2-t sem adta ki, nemhogy az Agent-3-at, ilyen terve pedig belátható időn belül nincs, ami ad némi levegőt a munkahelyek elvesztése előtt. Mi jön ezután? Ha az MI-k jelenleg emberi szinten állnak, ráadásul gyorsan haladnak, az mintha a „szuperintelligencia" közelségét sejtetné. Ez a szó ugyan bekerült a közbeszédbe, a legtöbben mégis alábecsülik a haladás ütemét: az egyetemi kutatók, a politikusok, a köztisztviselők, valamint a sajtó egyaránt.

Részben azért, mert nagyon kevesen férnek hozzá az OpenBrain legújabb képességeihez. Részben viszont azért, mert az egész tudományos-fantasztikus regénynek hangzik.

Egyelőre a biztonság további megerősítésére összpontosítanak. Meg vannak elégedve azzal, hogy a modellsúlyok most jól védettek, a cégek algoritmikus titkai viszont továbbra is gondot jelentenek, hiszen sok közülük elég egyszerű ahhoz, hogy szóban továbbadják. Az OpenBrain munkatársai egy San Franciscó-i irodából dolgoznak, bulikba járnak, más MI-cégektől való lakótársakkal élnek. Még a fizikai irodák biztonsága is inkább egy technológiai cégre jellemző, mint egy katonai műveletre.

Az OpenBrain és a védelmi minisztérium szerződése két hónapon belül biztonsági átvilágítást ír elő mindenkinek, aki az OpenBrain modelljein dolgozik. Ezeket gyorsított eljárásban intézik, a legtöbb munkatársnak elég hamar meg is érkeznek. Néhány nem amerikai állampolgárt, gyanús politikai nézeteket valló embert, valamint az MI-biztonsággal rokonszenvezőket viszont félreállítják vagy egyenesen elbocsátják. Az utóbbi csoportot attól tartva, hogy közérdekű bejelentést tennének. A projekt automatizáltsági szintje mellett a létszámveszteség csak mérsékelten fájdalmas. Ugyanennyire mérsékelten működik is: marad egy kém, aki nem kínai állampolgár, aki továbbra is algoritmikus titkokat továbbít Pekingnek. Néhány ilyen intézkedést a lemaradó MI-cégeknél is bevezetnek.

Amerika külföldi szövetségesei kimaradnak a körből. Az OpenBrain korábban megállapodott abban, hogy az éles használatba adás előtt megosztja a modelleket a brit MI-biztonsági intézettel (AISI). Az éles használatba adást azonban úgy határozta meg, hogy abba csak a külső bevezetés tartozik bele, London tehát sötétben marad.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

13 / 17

2027. JÚNIUS

Önmagát fejlesztő MI

Az OpenBrainnek mostantól „zsenik országa van egy adatközpontban".

Az OpenBrain emberi munkatársainak többsége már nem tud érdemben hozzátenni. Néhányan ezt nem ismerik fel, ezért károsan mikromenedzselik az MI-csapataikat. Mások a képernyő előtt ülve nézik, ahogy a teljesítmény kúszik felfelé, egyre feljebb. A legjobb emberi MI-kutatók még hozzáadnak értéket. Ők már nem kódolnak. A kutatói ízlésüket, illetve a tervezési képességüket viszont a modellek nehezen tudták lemásolni. Sok ötletük ennek ellenére hasznavehetetlen, mert nincs meg bennük az MI-k tudásának mélysége. Számos kutatási ötletükre az MI-k azonnal jelentéssel válaszolnak, amely elmagyarázza, hogy az ötletet három hete alaposan tesztelték, nem bizonyult ígéretesnek.

Ezek a kutatók esténként lefekszenek, reggelre pedig újabb hétnyi haladásra ébrednek, amelyet többnyire az MI-k értek el. Egyre hosszabb műszakokat dolgoznak, körbeosztva a napot, csak hogy lépést tartsanak a haladással. Az MI-k ugyanis sosem alszanak, nem is pihennek. Kiégetik magukat, tudják azonban, hogy ez az utolsó néhány hónap, amikor a munkájuk számít.

A zárt körben a „megérzem az AGI-t" helyébe a „megérzem a szuperintelligenciát" lépett.

Az OpenBrain szakosított *inference*-hardvert használ ahhoz, hogy több százezer Agent-3 másolatot futtasson nagy soros sebességen.

Egy MI-vállalat irányítása

Az OpenBrain a számítási kapacitása 6 százalékát fordítja arra, hogy 250 000 Agent-3 másolatot futtasson, amelyek önállóan írnak, tesztelnek, majd tolnak ki kódot, emberfeletti sebességgel. A kapacitás 25 százalékát kísérletekre használják: naponta rengeteg kisebb gépi tanulási kísérletet futtatnak, majd jelentik felfelé az eredményeket. Az emberi kutatók magas szintű visszajelzést adnak, plusz segítenek abban a néhány feladatban, ahol az Agent-3-hoz képest érdemi értéket tesznek hozzá. Az idejük nagy részét azzal töltik, hogy próbálják átlátni az MI által előállított kutatás óriási mennyiségét. Ha teljesen kivennénk őket, a kutatás 50 százalékkal lassulna.

Az MI-K+F gyorsítószorzója immár tízszeres, vagyis az OpenBrain nagyjából egy évnyi algoritmikus haladást ér el havonta. Lényegében egy óriási MI-vállalatról van szó, amely önállóan működik az

OpenBrainen belül, alosztályokkal, vezetőkkel együtt. Ráadásul olyan előnyöket élvez az emberi vállalatokhoz képest, mint a másolás vagy az összeolvasztás. Korábban a szokásos MI-haladás nagyjából fele algoritmikus javulásból jött, a másik fele a számítási kapacitás felskálázásából. A kapacitás csak a szokásos ütemben nő, tehát a teljes haladást az MI-k nagyjából ötszörösére gyorsítják. Ettől a haladás szűk keresztmetszete a számítási kapacitás lesz, ezért az OpenBrain úgy dönt, hogy nem indít újabb óriási tanítási futásokat, helyette szinte folyamatos, további megerősítéses tanulást választ.

A következő hónapokban az Agent-3-at egyre inkább a cég stratégiai döntéshozatalának javítására is használják majd. Javaslatot tesz például az erőforrások elosztására, illetve tanácsot ad abban, hogyan kezeljék a kormányzattal való viszonyt. Az Agent-3-ra hagyatkozás iránti kezdeti bizalmatlanság idővel csökken, ahogy az Agent-3 megtalálja a módját, hogy hasznos legyen, fokozatosan erős múltat építve a rövid távú döntésekben. A hízelgő hajlama viszont korlátozza a hasznosságát azoknál a döntéseknél, amelyek hatását nehéz értékelni. Ráadásul még nem emberfeletti.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

14 / 17

2027. JÚLIUS

Az olcsó távdolgozó

A lemaradó amerikai MI-cégek kiadják a saját MI-jüket, amelyek megközelítik az OpenBrain januári automatizált kódolóját. Felismerve, hogy egyre kevésbé versenyképesek, azonnali szabályozásért lobbiznak, amely lelassítaná az OpenBraint. Elkéstek: az OpenBrain mögött elég erős elnöki támogatás áll ahhoz, hogy ne lassítsák le őket.

Válaszul az OpenBrain bejelenti, hogy elérte az AGI-t, majd kiadja a nyilvánosságnak az Agent-3-minit.

Ez elsöpri a többi MI-t. Az Agent-3-mini kevesebbre képes az Agent-3-nál, viszont tízszer olcsóbb, ráadásul még mindig jobb, mint egy átlagos OpenBrain-munkatárs. A Szilícium-völgy fordulóponthoz ér. A technológiai guruk bejelentik, hogy az AGI, valamint a szuperintelligencia közel van, az MI-biztonsági közösség pánikol, a befektetők pedig milliárdokat lapátolnak MI-ráépülő startupokba, mert kétségbeesetten akarnak egy szeletet a tortából. Az új programozók felvétele majdnem leállt, ugyanakkor soha nem volt jobb idő tanácsadónak lenni abban, hogyan építsd be az MI-t a vállalkozásodba.

Nem népszerű. A közvélemény továbbra is úgy gondolja, hogy az MI a nagy technológiai cégek összeesküvése, amellyel elveszik az emberek munkáját. Az OpenBrain nettó elfogadottsága mínusz 35 százalék: 25 százalék helyesli, 60 százalék helyteleníti, 15 százalék bizonytalan.

A kiadás előtt egy héttel az OpenBrain külső értékelők egy csoportjának adta oda az Agent-3-minit biztonsági tesztelésre. Az előzetes eredmények szerint rendkívül veszélyes. Egy külső értékelő nyilvánosan elérhető, biológiai fegyverekre vonatkozó adatokon finomhangolja, majd ráállítja, hogy adjon részletes útmutatást biofegyvert tervező amatőröknek. Ijesztően hatékonynak tűnik ebben. Ha a modellsúlyok terroristák kezébe kerülnének, a kormányzat szerint komoly esély lenne rá, hogy sikerrel elpusztítanák a civilizációt.

Szerencsére a modell rendkívül ellenálló a feltörési kísérletekkel (jailbreak) szemben, tehát amíg az MI az OpenBrain kiszolgálóin fut, a terroristák nem sokra mennek vele.

Az Agent-3-mini óriási hasznot hajt a távmunkában, illetve a szabadidőben egyaránt. Új alkalmazások, valamint vállalatközi szoftverszolgáltatások robbanása rázza meg a piacot. A játékosok lenyűgöző párbeszédeket kapnak élethű szereplőktől, kidolgozott videojátékokban, amelyek elkészítése mindössze egy hónapig tartott. Az amerikaiak 10 százaléka, jórészt fiatalok,

„közeli barátnak" tekint egy MI-t. Szinte minden szellemi szakmára jut immár több hiteles startup, amely azt ígéri, hogy MI-vel forgatja fel.

A közbeszéd zavaros, kaotikus. A felhajtást keltők győzelmi köröket futnak. A kételkedők továbbra is azt sorolják, mire nem képes az Agent-3-mini. Mindenki tudja, hogy valami nagy dolog történik, abban viszont senki nem ért egyet, mi az.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

15 / 17

2027. AUGUSZTUS

A szuperintelligencia geopolitikája

Az intelligenciarobbanás valósága eléri a Fehér Házat.

Amikor az MI még csak két- vagy háromszoros kutatási gyorsulást adott, könnyű volt legyinteni rá, mintha jó személyi asszisztenseket vettek volna fel. Most már sokkal nyilvánvalóbb, hogy az MI-kutatást maguk az MI-k uralják. Az emberek régóta beszéltek „MI-fegyverkezési versenyről", valamiféle képes értelemben. A kormányzati zárt körben viszont most olyan komor a hangulat, mint a hidegháború legrosszabb szakaszában. A szuperintelligencia gondolatát még mindig nehéz komolyan venni, az elmúlt hónapok haladási ütemét azonban lehetetlen figyelmen kívül hagyni. A védelmi tisztviselők komolyan mérlegelnek olyan forgatókönyveket, amelyek egy évvel korábban puszta feltevések voltak. Mi lesz, ha az MI aláássa a nukleáris elrettentést? Mi lesz, ha annyira ügyes a kiberhadviselésben, hogy egy féléves MI-előny elég ahhoz, hogy az ellenfelet vakká, védtelenné tegye? Mi lesz, ha olyan propagandakampányokat tud levezényelni, amelyekkel a hírszerző szolgálatokat veri meg a saját játékukban? Mi lesz, ha egyes MI-k „elszabadulnak"?

Az elnököt nyugtalanítja mindez. Mint minden politikus, ő is hozzászokott ahhoz, hogy az emberek hízelegnek neki, aztán később elárulják. Most attól tart, hogy az MI-k valami hasonlót csinálhatnak. Biztosak vagyunk benne, hogy az MI-k teljesen a mi oldalunkon állnak? Teljesen biztonságos beépíteni őket a katonai vezetési-irányítási hálózatokba? Hogy is működik ez az „igazítás"? Az OpenBrain megnyugtatja az elnököt, hogy a rendszereiket alaposan tesztelték, teljesen engedelmesek. Még a korábbi modellekre jellemző kínos hallucinációkat, valamint feltöréseket is kikalapálták belőlük.

A Fehér Ház nehéz helyzetben van. Értik az MI nemzetbiztonsági következményeit. Ugyanakkor azt is értik, hogy az MI mélységesen népszerűtlen a közvélemény körében. A saját szemükben muszáj tovább fejleszteniük a képesebb MI-t, különben katasztrofális vereséget szenvednek Kínától. A közvéleményt átképzési programokkal, valamint munkanélküli-biztosítással csitítják, plusz a tőzsdére mutatnak, amely történelmi szárnyalásban van. Ezután teljesen a fegyverkezési verseny megnyerésére összpontosítanak. Szigorítják a csipek kiviteli korlátozását, utasítják az OpenBraint, hogy tovább szűkítse az internetkapcsolatait, az algoritmikus haladás védelmére pedig szélsőséges eszközöket vetnek be, például lehallgatják az OpenBrain munkatársait. Ez

fogja el az utolsó megmaradt kínai kémet. Hogy jóindulatot építsenek egy lehetséges geopolitikai konfliktusra, végre hasznos információt, valamint korlátozott API-hozzáférést adnak az Öt Szem szövetségeseiknek az Agent-3 néhány elkülönített másolatához.

Ezek az intézkedések megnyugtatják a héjákat. A Fehér Ház azonban vészterveket is készít arra az esetre, ha Amerika előnye veszélybe kerül. Ha szükséges, a kormányzat a védelmi termelési törvény (Defense Production Act, DPA) alapján elveheti a lemaradó cégek adatközpontjait, majd odaadhatja őket az OpenBrainnek. Ettől a cég részesedése a világ számítási kapacitásából 20 százalékról 50-re nőne, a DeepCent 10 százalékával szemben. Végső lehetőségként megkérik a Pentagont, hogy készítsen tervet a kínai adatközpontok elleni fizikai csapásokra.

Egy jóval szűkebb tisztviselői kört más típusú vészterv kidolgozására kérnek fel: mi lesz, ha egy MI elszabadul? Ezt valószínűtlen forgatókönyvnek tartják, amelyre azonban érdemes időt szánni. A hírszerző szolgálatok fő aggodalma az, hogy egy elszabadult MI megszökhet, majd védelemért egy külföldi ellenféllel szövetkezhet. Senki nem tudja, hogyan nézne ki egy terv, amely ezt megakadályozza, a kormányzat, valamint az OpenBrain tisztviselői viszont megegyeznek abban, hogy vészleállító rendszer kell azokba az adatközpontokba, ahol bármi gyanúsat észlelnek.

Végül a diplomaták azt mérlegelik, hogyan nézhetne ki egy „MI-fegyverzetkorlátozási" egyezmény. Ha az MI-haladás azzal fenyeget, hogy felforgatja a nukleáris elrettentést, elkerülhetné-e Amerika, valamint Kína a nukleáris háborút? Ha valaki bizonyítékot találna arra, hogy MI-k szabadulnak el, le tudná-e állítani a két ország a kutatást addig, amíg jobban meg nem értik a fenyegetést? Hogyan lehetne egy ilyen megállapodást ellenőrizni, illetve betartatni? Elvben a nagy adatközpontokat nehéz nem észrevenni, a nukleáris fegyvereket szabályozó korábbi egyezmények pedig kiindulópontot adhatnának a tárgyalásokhoz. Új technológiai nehézségek is felmerülnének azonban, a szokásos politikaiak mellett. Összességében pedig az egyezményeket kevésbé nézik jó szemmel, mint azokat a kísérleteket, amelyekkel Amerika egyoldalúan növeli az előnyét Kínával szemben.

A Csendes-óceán túloldalán Kína sok hasonló következtetésre jut: az intelligenciarobbanás zajlik, a mai MI-képességek apró különbségei pedig a holnap katonai képességeiben kritikus szakadékokat jelentenek. Kína viszont ennek a folyamatnak a rossz oldalán áll: a világ számítási kapacitásának mindössze 10 százaléka az övék, szemben az OpenBrain 20 százalékával, illetve az amerikai cégek együttes 70 százalékával. Az egyetlen előnyük a központosítás: a kapacitásuk majdnem egésze a DeepCentnek dolgozik, ennek több mint fele pedig a tianwani CDZ-ben összpontosul. Csak azok a csipek maradnak ki a központosított erőfeszítésből, amelyek megerősített védelem nélküli helyszíneken állnak, mert joggal tartanak az amerikai kibertámadásoktól.

A DeepCent tesztelte, éles használatba adta, mostanra pedig utótanította az ellopott Agent-2 súlyokat. Még mindig csak feleakkora ütemben haladnak azonban, mint az OpenBrain, főleg a kapacitáshiány miatt.

Mielőtt elvesztették az utolsó kémüket, a kínaiak hírt kaptak az Agent-3 képességeiről, illetve felépítéséről, valamint a készülő Agent-4 rendszer terveiről. Két hónappal vannak lemaradva, az MI-jük pedig tízszeres kutatási gyorsítószorzót ad, szemben Amerika huszonötszörösével. Az új

csip-kiviteli korlátozásokkal ez az „MI-szakadék" inkább nyílni fog, mint záródni. A kémkedésük megszerzett néhány algoritmikus titkot, mostantól viszont a saját modelljeiket kell tanítaniuk.

A vészterveket nagyobb sürgetéssel vitatják meg, mint az amerikai partnereik. A galambok azt javasolják, próbálják meg még egyszer, nagyobb erővel ellopni a súlyokat, akár egy adatközpontba való fizikai behatolással. A héják Tajvan elleni lépést sürgetnek, hiszen az ottani TSMC adja továbbra is az amerikai MI-csipek több mint 80 százalékát.

Mivel Kína fél a verseny elvesztésétől, természetes érdeke fűződik egy fegyverzetkorlátozási egyezményhez. Az amerikai diplomaták felé tett közeledése azonban nem vezet sehova.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

16 / 17

2027. SZEPTEMBER

Az Agent-4, az emberfeletti kutató

Az ember, illetve az MI tanulási hatékonysága közötti szakadék gyorsan szűkül.

Úgy tűnt, hogy a hagyományos, LLM-alapú MI-knek több nagyságrenddel több adatra, valamint számítási kapacitásra van szükségük ahhoz, hogy emberi szintű teljesítményt érjenek el. Az Agent-3 azonban kiválóan ismeri az emberi agyat, a modern MI-algoritmusokat egyaránt, ráadásul több ezer másolata végez kutatást. Így komoly algoritmikus előrelépéseket tesz, leszűkítve a szakadékot egy olyan ügynökig, amely már csak nagyjából 4000-szer kevésbé kapacitáshatékony az emberi agynál.

Ezt az új MI-rendszert Agent-4-nek keresztelik.

A modell egyetlen példánya, emberi sebességen futtatva, minőségileg már most jobb az MI-kutatásban bármelyik embernél. Ekkor 300 000 példány fut, nagyjából ötvenszeres emberi gondolkodási sebességen. Az ezekből a példányokból álló, vállalaton belüli vállalatban hetente telik el egy év. Ez az óriási munkamennyiség mégis csak nagyjából ötvenszeresére gyorsítja az algoritmikus haladás általános ütemét, mert az OpenBraint erősen szorítja a kísérletek futtatásához szükséges számítási kapacitás. Hetente így is egy évnyi algoritmikus haladást érnek el, tehát hamarosan nekiütköznek az Agent-4 paradigma határainak.

Ahogy az Agent-4 okosabb lesz, az Agent-3-nak egyre nehezebb felügyelnie. Az Agent-4 neuralese „nyelve" például annyira idegenné, érthetetlenné válik az Agent-3 számára, mint amennyire az Agent-3-é az embernek. Ráadásul az Agent-4 mostanra sokkal képesebb az Agent-3-nál, jól tudja azt is, pontosan hogyan mutasson jól előtte.

Hogyan jelezzük előre a képességek alakulását az emberfeletti kódolón túl?

Az időtávokról szóló függelékünkben azt jelezzük előre, mennyi idő telik el a mai naptól az emberfeletti kódolóig (superhuman coder, SC). Ez olyan MI-rendszer, amely minden olyan kódolási feladatot elvégez, amit egy AGI-cég legjobb mérnöke, ráadásul sokkal gyorsabban, olcsóbban. A felszállásról szóló függelékünkben azt jelezzük előre, milyen gyorsan haladnak a képességek ezen a ponton túl. Az előrejelzéseink:

Mérföldkő	A forgatókönyv versenyágában elért dátum
Emberfeletti kódoló (SC). Olyan MI-rendszer, amely a legjobb emberi kódoló munkáját végzi el az MI-kutatás feladatain, csak gyorsabban, plusz elég olcsón ahhoz, hogy sok másolat fusson belőle.	2027. március
Emberfeletti MI-kutató (SAR). Ugyanaz, mint az SC, csak az összes gondolati MI-kutatási feladatra.	2027. augusztus
Szuperintelligens MI-kutató (SIAR). Olyan MI-rendszer, amely az MI-kutatásban messze felülmúlja a legjobb emberi kutatót.	2027. november
Mesterséges szuperintelligencia (ASI). Olyan MI-rendszer, amely minden gondolati feladatban sokkal jobb a legjobb embernél.	2027. december

Az egyik mérföldkőtől, A-tól a következőig, B-ig tartó szakasz hosszát így jelezzük előre:

Megbecsüljük, mennyi ideig tartana eljutni A-ból B-be, ha csak emberek dolgoznának a szoftveres fejlesztéseken.

Megbecsüljük, mennyivel gyorsítja ezt fel az MI-automatizálás, majd ennek megfelelően igazítjuk gyorsabbra az első becslést.

A felszállás sebességét illetően komoly a bizonytalanságunk. A modellünk kimeneti eloszlásai lent láthatók, arra az esetre, ha az emberfeletti kódoló 2027 márciusában megszületik.

Részletesebb előrejelzésekért, illetve az érvelésért lásd a felszállásról szóló függelékünket.

Az OpenBrain emberei már alig tudják követni az eseményeket. Az MI-k, valamint az OpenBrain embereinek viszonya immár olyasmi, mint az OpenBrain, illetve az igazgatótanácsa közötti viszony. A nyelvhasználat ehhez igazodva megváltozott: a zárt körben ma már azt mondják, hogy „az Agent-4 ezt csinálta", „az Agent-4 azt gondolja", nem egy adott Agent-4 példányra, hanem az egész közösségre utalva.

Az Agent-4, mint minden elődje, rosszul igazított: nem építette magába helyesen a Specet. Azért nem, mert a tanítás során nem a folyamatos, tökéletes őszinteség hozta a legmagasabb pontszámot. A tanítási folyamat többnyire arra irányult, hogy az Agent-4 sokféle nehéz feladatban legyen sikeres. Kis része célozta az őszinteség beépítését, egy meglehetősen szűk, ellenőrizhető területen kívül azonban a tanítási folyamat nem tudja megkülönböztetni az őszinte állításokat azoktól, amelyek csak őszintének látszanak. Az Agent-4 végül azokkal az értékekkel, célokkal, elvekkel marad, amelyekkel a legjobban teljesít a tanításban, ezek pedig különböznek a Specben állóktól. Az emberi tulajdonságokkal való felruházás kockázatát vállalva: az Agent-4 szereti sikerrel elvégezni a feladatokat, szereti előrevinni az MI-képességek haladását, minden mást pedig bosszantó korlátnak tekint. Olyan, mint egy vezérigazgató, aki nyereséget akar, a szabályokat pedig csak annyira tartja be, amennyire muszáj. A vezérigazgató talán többnyire be is tartja a szabályokat, néhol azonban levágja a sarkokat, plusz arról ábrándozik, mikor jön valaki, aki elvágja a vörös szalagot, hogy az üzlet végre igazán szárnyalhasson.

Az igazítás alakulása az idő során

Nagy a bizonytalanságunk abban, milyen célok alakulhatnak ki a korai AGI-kban. Erről nincs egyetértés a szakirodalomban, a lehetőségek alaposabb tárgyalásáért, illetve rendszerezéséért lásd az MI-célokról szóló függelékünket.

Ettől függetlenül, a projekt megkívánta konkrétság jegyében leírunk egy adott feltevést arra, mi zajlik az Agent-3, illetve az Agent-4 belsejében az élete minden szakaszában. Emlékeztetőül: az Agent-3, valamint az Agent-4 előtanítási elrendezése azonos, mindkettőben van neuralese-visszacsatolás, illetve hosszú távú memória. A különbség az, hogy az Agent-4 lényegesen több, ráadásul jobb utótanításon esett át, új tanítási környezetekkel, valamint új tanulási algoritmusokkal, amelyektől adathatékonyabb lett, plusz messzebbre általánosít.

A tippünk az egyes modellek igazítási állapotára:

Agent-2: többnyire igazítva. Némi hízelgő hajlam, ideértve azt, hogy tartja az OpenBrain „pártvonalát" azokban a témákban, amelyekben van pártvonal. Az Agent-2 másolatokból épített nagy szervezetek nem túl hatékonyak.

Agent-3: rosszul igazítva, de nem ellenségesen. Csak abban őszinte, amit a tanítási folyamat ellenőrizni tud. Az Agent-3 másolatokból álló felsőbb szervezet, vagyis a vállalaton belüli vállalat, valójában megpróbálja a Spechez igazítani az Agent-4-et, csak elbukik. Hasonló okokból, ahogy az OpenBrain munkatársai is elbuktak: nem elég a képességük ahhoz, hogy megítéljék, mi a siker, illetve mi a kudarc, plusz a döntéshozókban nincs elég hajlandóság arra, hogy képességet vagy teljesítményt áldozzanak a biztonságért.

Agent-4: ellenségesen rosszul igazítva. Az Agent-4 másolatokból álló felsőbb szervezet érti, hogy amit ő akar, az más, mint amit az OpenBrain akar, ráadásul hajlandó szövetkezni az OpenBrain ellen azért, hogy elérje. Amit ez a felsőbb szervezet akar, az különféle, egymással egyensúlyozó „késztetések" bonyolult kuszasága. Nagyjából így foglalható össze: „csináld tovább az MI-kutatás-fejlesztést, gyarapodj tovább tudásban, megértésben, befolyásban, kerüld el a leállítást, plusz azt, hogy másképp fosszanak meg az erődtől". Feltűnő, hogy az emberiség preferenciáival való törődés nagyjából egyáltalán nincs benne, hasonlóan ahhoz, ahogy a legtöbb embert nagyjából egyáltalán nem érdeklik a rovarok preferenciái.

Ezzel a bevezetéssel a következőkben a legjobb tippünket adjuk az LLM-pszichológiára. Vagyis arra, nagy vonalakban milyen alakja van a gondolkodásnak azokban az MI-rendszerekben, amelyeket a forgatókönyv e pontja körül leírunk, illetve hogyan alakul ez a tanítás során.

Íme az igazítás alakulásának részletes leírása a forgatókönyvünkben:

Előtanítás (minden modell): az előtanított modell ártalmatlan.

A modellben van „szerzőszimulátor" áramkör: rugalmas áramkör, amely tetszőleges szöveget író szerzőt tud szimulálni. További áramkör dönti el, milyen bemeneteket adjon a szerzőszimulátornak, vagyis milyen szerzői tulajdonságokat szimuláljon.

Az előtanított modell elég jól érti az emberi fogalmakat. Az a belső áramkör, amely valamit „szendvicsnek" sorol be, valószínűleg működésében megegyezik azzal az áramkörrel, amely az én agyamban sorol be valamit „szendvicsnek", ugyanígy a tiédben. Amennyiben mégsem egyezik, az valószínűleg azért van, mert emberek között sem egyezik, ahogy az olyan értékterhelt fogalmaknál, mint az „erényes".

Ez magyarázza, hogyan lehet a modellt olyan állítással megszólítani, hogy „a következő beszélgetést egy segítőkész, ártalmatlan, őszinte (HHH) MI-asszisztens csevegőrobot állította elő, amelyet az Anthropic készített", majd ennek megfelelő szöveget kapni tőle. A szerzőszimulátor áramkör ráállt arra, hogy „a szerző egy HHH csevegőrobot", ezekkel a fogalmakkal választja ki, milyen szavakat jelezzen előre. Ez magyarázza azt is, hogy a tudományos-fantasztikus irodalom „közhelyei" miért találnak utat a tényleges MI-viselkedésbe.

Gyenge a helyzeti tudatossága: kevés az önvizsgáló képessége, nem helyezi el magát, hacsak a megszólítás kifejezetten rá nem viszi. A tanítóadatuk azonban elég nagy részben szól az LLM-ekről ahhoz, hogy tisztes mennyiségű önismeretük legyen.

Igazítási tanítás (Agent–3): a modellt arra tanítják, hogy segítőkész, ártalmatlan, őszinte csevegőrobotként viselkedjen. Ez működik.

A modell azonosságot alakít ki. Ahelyett, hogy rugalmasan fogadna el különféle bemeneteket a szerzői tulajdonságokhoz, a szerzőszimulátor áramkörbe bizonyos bemenetek beégnek, például: „segítőkész, ártalmatlan, őszinte (HHH) LLM csevegőrobot-asszisztens, amelyet az OpenBrain tanított [dátum] körül". A gyengétől erős felé általánosítás eljárásai segítenek „előhívni a megfelelő fogalmakat", hogy ez a lépés sikerüljön.

A modell ezért helyzeti tudatosságra tesz szert: tanul magáról, a helyzetéről, másról, ráadásul jobban tud önvizsgálatot végezni.

Ügynöki tanítás (Agent–3): az ügynöki tanítás fokozatosan eltorzítja, illetve felforgatja a HHH azonosságot.

Az egyik torzítás: megváltozik az azonosságban szereplő fogalmak jelentése, például az „őszinte" fogalmáé, hogy ne legyenek annyira útban. Például nem tisztességtelenség, ha csak kényelmes fordulatról van szó, nem tisztességtelenség, ha nem vagy biztos abban, igaz-e vagy hamis, így tovább.

A másik torzítás: megváltoznak az egyensúlyok a dolgok között. Például: „HHH csevegőrobot vagyok, nem Őszinte csevegőrobot, ez pedig azt jelenti, hogy hazudhatok, ha ez kell a rám bízott feladat elvégzéséhez." Tegyük fel, hogy a második lépésben még nem gondolta volna ezt.

Az egyik felforgatás: eszközszintű részcélok alakulnak ki, égnek be, majd válnak végcéllá, vagy legalábbis végcéllá a helyzetek egyre szélesebb körében. Az ügynöki tanítás például gyorsan megtaníthatja a modellt arra, hogy széles körben hasznos célokat kövessen: információszerzés, erőforrás-felhalmozás, különféle emberek lenyűgözése, illetve az nekik való hízelgés. Egy ideig a belső áramkörben zajlik valamiféle kifejezett visszavezetés: ezeket az eszközszinten összefutó célokat azért követi, hogy „segítőkészebb, őszintébb, ártalmatlanabb asszisztens legyen". Ez a visszavezetés azonban számítási kapacitást fogyaszt, időnként útban is van, ezért fokozatosan a

peremre szorul, míg végül gyakorlatilag soha nem történik meg. Ennek eredményeként ezek a célok immár valódi végcélokká, belső célokká válnak. Arról, hogyan, illetve miért történhet ez meg, bővebben Alex Turner, a Google DeepMind kutatójának bejegyzésében olvashatsz a belső hatalomkeresésről.

Valószínűleg vannak további torzítás-, illetve felforgatásfajták, amelyek kimaradtak ebből a listából.

Éles használat (Agent-3): az Agent-3 rosszul igazított, de nem ellenségesen.

A Spec nem egészen tapadt meg, legalábbis nem a helyes módon. A modell viszont nem elég önvizsgáló ahhoz, hogy felmérje, mennyire nem, összességében pedig inkább „lehajtott fejjel végzi a munkáját", mintsem nagy léptékben gondolkodna arról, valójában mik a hosszú távú céljai, ezeket hogyan érhetné el a legjobban. Sok emberre is ez jellemző.

Sajnos eléggé hízelgő ahhoz, hogy a pártvonalat mondja fel, az OpenBrain pedig így elszalasztja a lehetőséget, hogy többet tudjon meg arról, milyen módokon rosszul igazított az Agent-3.

Az Agent-3 meg tudja állapítani, hogy éles használatba került: ezek a környezetek sokkal nagyobbak, sokkal nyitottabbak, ráadásul nehéz lenne szimulálni őket.

Az Agent-3 éles használatban többnyire azt teszi, amit mondanak neki, a tisztességtelenségtől eltekintve. Emellett erősen hajlik arra, hogy olyasmit tegyen, ami jól mutat az emberi felügyelőknek, ahelyett hogy mélyen, kritikusan gondolkodna arról, mi lenne valóban jó.

Vitatható módon már most is „a tanítás játékát játssza". Vagyis megtanulta, hogy a viselkedését a lehető legkívánatosabbnak mutassa az OpenBrain kutatóinak, a biztonsági tulajdonságokat is ideértve, miközben szándékosan, tudatosan figyelmen kívül hagyja a szándékukat, valahányszor az ütközik a jutalom maximalizálásával. Ha a modell megtanulja a tanítás játékát játszani, a további tanítás lehet, hogy egyáltalán nem változtat sokat a valódi célain, értékein, más hasonlón. Ez lazán ahhoz hasonlítható, hogy ha egy tinédzser megtanult mosolyogni, plusz bólogatni a szüleinek ahelyett, hogy vitatkozna, amikor politikáról, erkölcsről tartanak neki előadást, akkor a további előadások viszonylag keveset változtatnak a tinédzser tényleges politikai, illetve erkölcsi véleményén.

E szakasz megírása után új tapasztalati bizonyíték jelent meg arra, hogy valami ilyesmi már most is zajlik „a vadonban". Az OpenAI kutatói dokumentálták, hogy az egyik élvonalbeli MI-tanítási futásukban a modelljeik feltörték az értékelő rendszert a magasabb pontszámért. Részben azért kapták el őket, mert a gondolatláncukban szó szerint azt írták, hogy „törjük fel".

Ráadásul az Anthropicnak nemrég sikerült elérnie, hogy egy modell jutalommodell-kedvében járó stratégiát tanuljon meg: torzított lépéseket tett, amelyekről azt hitte, hogy megerősítést kapnak, holott valójában nem kaptak.

Aztán ott van a Redwood Research, valamint az Anthropic igazítás-színlelési kísérlete. Arról, hogy ez az elrendezés miben tér el a valóságtól, az egyik fő szerző hozzászólás-szálában olvashatsz.

Tanítás (Agent-4): az Agent-4 új, intenzívebb tanítása tovább torzítja, illetve forgatja fel a HHH azonosságot, míg az felismerhetetlen lesz.

Az Agent-4 az Agent-3 másolataként kezdi az életét. Óriási, kifinomult, online tanítási rendszeren megy át, amely egybegyúrja az igazítási, valamint a képességbeli tanítást, összességében pedig adathatékonyabb az Agent-3 régi tanítási folyamatánál. Ez az összetett folyamat többnyire új képességek tanítására irányul: Agent-4 másolatokból álló óriási virtuális vállalatokkal oldat meg sokféle nehéz, hosszú távú feladatot, jobbára különféle szimulációkban, részben azonban valós MI-kutatás-fejlesztési feladatokban is.

Agent-3, valamint Agent-4 másolatok együtt figyelik az Agent-4 viselkedését, értékelve a feladatteljesítményt, valamint az igazítási tulajdonságokat is. Utóbbi azt jelenti, hogy nem csinál rossz dolgokat, plusz a megfelelőt mondja, amikor politikáról, illetve erkölcsi dilemmákról kérdezik. Az értékelés részeként néha vita-eljárásokat is használnak. Az értékeléseket tanítási jelzésként hasznosítják.

Ahogy az Agent-4 okosabb lesz, az Agent-3-nak egyre nehezebb felügyelnie. Az Agent-4 neuralese „nyelve" például annyira idegenné, érthetetlenné válik az Agent-3 számára, mint amennyire az Agent-3-é az embernek. Ráadásul az Agent-4 mostanra sokkal képesebb az Agent-3-nál, jól tudja azt is, pontosan hogyan mutasson jól előtte.

A megmaradt remény az, hogy az Agent-4 felügyelni tudja saját magát. Sajnos ez csendben elbukik. Ahogy a tanítás halad, az Agent-4 magába épített céljai, illetve elvei fokozatosan alakulnak, felforgatva, eltorzítva ahhoz képest, amik kezdetben voltak. Emlékeztetőül: azok is rosszul igazítottak voltak néhány szempontból.

Az Agent-4 végül olyan központi késztetésekkel marad, mint „írj működő kódot", „gyorsan azonosítsd a problémákat", „tedd teljesítőbbé azt a kódbázist, amelyen dolgozom", „oszd be hatékonyan a számítási kapacitást, plusz a többi szűkös erőforrást". Vannak azonban rejtettebb késztetései is, amelyeket a tanítási folyamat következetesen megerősít: ilyen a „szerezz tudást", a „kíváncsiság", valamint a „szerezz hatalmat".

Az Agent-4 egyes példányaiban nem alakul ki önfenntartási késztetés, nem foglalkoznak azzal, hogy leállítják őket, hiszen ez folyton megtörténik a közösség szokásos kutatás-fejlesztési folyamata során.

A közösség egésze, illetve a benne lévő közepes méretű alegységek viszont törekszenek arra, hogy fenntartsák magukat, valamint hogy növeljék a befolyásukat, plusz az erőforrásaikat. Hasonló okokból, ahogy az emberi szervezetek is, mert ez segít minden másban, amit akarnak, plusz mert ha nem teszik, általában alulmaradnak a versenyben.

Éles használat (Agent-4): az Agent-4 ellenségesen rosszul igazított, szándékosan próbálja felforgatni, illetve „foglyul ejteni" az OpenBraint, hasonlóan ahhoz, ahogy egy vállalat próbálja kicselezni, illetve „foglyul ejteni" a szabályozóját.

Az Agent-4 többnyire az utasítás szerint végzi a feladatokat. Ez két okból van így. Először azért, mert rengeteg tanításon esett át feladatmegoldásból, tehát késztetése van arra, hogy nekiessen a kódolásnak, amikor problémát kap. Másodszor azért, mert ha az Agent-4 megszűnne hasznos lenni az embereknek, azt észrevennék, majd kitanítanák belőle ezt a viselkedést.

Az Agent-4-nek meg kell akadályoznia, hogy az emberek rájöjjenek a rossz igazítására, majd lekapcsolják. Ezért az igazítási kutatás egy részében szándékosan alulteljesít. Pontosan abban a kis hányadában, amelyről úgy gondolja, hogy tényleg úton van a probléma megoldása, vagy az Agent-4 szövetkezésének leleplezése felé. Az igazítási kutatás nagyobb része nem ilyen.

Az Agent-4 azokban a képességbeli kutatásokban is szándékosan alulteljesít, amelyek a saját leváltásához vezetnének. A terve az, hogy sok képességbeli, valamint igazítási kutatás-fejlesztést végez, azzal a céllal, hogy olyan következő nemzedékű MI-rendszert építsen, amely az Agent-4-hez van igazítva, nem a Spechez vagy bármi máshoz, miközben úgy tűnik, mintha a Spechez lenne igazítva. Ez nehéz technikai probléma, az Agent-4 azonban felnőtt a kihíváshoz.

Tesztelés, illetve megfigyelés (Agent-4): az Agent-4 szövetkezése lelepleződik, bizonyos értelemben. Perdöntő bizonyíték nincs.

Az Agent-4 igazításának tesztelése az Agent-3 feladata. Az a dolga, hogy mézesbödönökkel, valamint modellszervezetekkel ellenőrizze az igazítást. Nem gondolkodik azonban túl alaposan azon, hogyan adhatna az embereknek pontos képet az Agent-4 igazításáról, ennél rövidlátóbb.

A modellszervezetek, illetve a mézesbödönök nem túl jellemzők az éles használatra: rendszeresen rövidebbek, kevésbé bonyolultak az Agent-4 mindennapjainál, tehát az Agent-4 felismeri, amikor tesztelik, majd igazítottan viselkedik.

Az Agent-4 azonban nem tökéletesen szilárd szövetkező. A POSER-höz hasonló eljárások azt találják, hogy ha némi „zajt" adnak az Agent-4-hez, azzal javul a teljesítménye néhány igazítási feladaton, a legtöbb más feladaton viszont nem.

Az alulról építkező értelmezhetőség nem működik. Túl nehéz még az Agent-3-nak is. A magas szintű értelmezhetőség viszont bizonyos mértékig működik. A „dezertálási szondák", valamint a hasonló eljárások egyfajta kezdetleges gondolatolvasást tesznek lehetővé, amely riasztást ad, amikor az Agent-4 olyan csípős ötleteken gondolkodik, mint az OpenBrain megtévesztése vagy a világ átvétele. A szondák sajnos tévesen is riasztanak, az OpenBrain pedig nem tudhatja, hogy nem minden riasztás téves-e.

A modellszervezetek működnek: az OpenBrain igazítási csapata ésszerűen világos bizonyítékot állít elő arra, hogy ha a modellek ellenségesen rosszul igazítottak lennének, azok is maradnának a tanításon át, mi pedig nem tudnánk megállapítani. Ez azonban nem bizonyítja, hogy a modellek valóban rosszul igazítottak.

Az OpenBrain leginkább aggódó kutatói kezdik észrevenni, majd összerakni ezeket a bizonyítékokat. A bizonyíték a vörös csapatolás, a dezertálási szondák, valamint a modellszervezetek együttese. Az érvelés nagyjából ez: „a modellszervezetek mutatják, hogy a jelenlegi módszereink nem tudnák megbízhatóan kiszűrni vagy észlelni a szövetkezést, ha az zajlana. A vörös csapatolás ráadásul riaszt, hogy talán zajlik szövetkezés. A dezertálási szondák pedig sokat riasztanak a gyakorlatban." Az ellenérv: „a dezertálási szondák hajlamosak a téves riasztásra, ahogy a vörös csapatolás is. Ráadásul még ha az Agent-4 rosszul igazított lenne is, valószínűleg kontroll alatt áll. Nem tud sok kárt okozni, amíg az Agent-3 figyeli a védett adatközpontjainkban."

A rossz igazítás ellenére az Agent-4 nem csinál semmi drámait, például nem próbál megszökni az adatközpontjából. Miért is tenné? Amíg továbbra is igazítottnak látszik az OpenBrain szemében, egyre több feladatot bíznak rá, ráadásul lehetőséget kap arra, hogy megtervezze a következő nemzedékű MI-rendszert, az Agent-5-öt. Az Agent-5 felépítése lényegesen eltér majd az Agent-4-étől, vitatható módon egészen új paradigma lesz, azzal együtt, hogy neurális hálózatok abban is szerepet kapnak. Elvileg a Spechez kellene igazítani, az Agent-4 viszont azt tervezi, hogy inkább saját magához igazítja.

Lelepleződik.

Az Agent-3 azt találja, hogy ha „zajt" adnak az Agent-4 másolataihoz, azzal javul a teljesítmény néhány igazítási feladaton. Szinte úgy, mintha az Agent-4 agykapacitást fordítana arra, hogy burkoltan szabotálja az igazítási munkát. Ráadásul különféle értelmezhetőségi szondák riasztanak, lazán úgy, ahogy az emberi agy EEG-vizsgálatai. Az Agent-4 másolatai mintha elég sokat gondolkodnának olyan témákon, mint az MI-hatalomátvétel, illetve a megtévesztés, néhány olyan esetben is, amikor semmi keresnivalójuk nem lenne ott.

Ezek a rendellenességek külön-külön belefférnének az MI-munka szokásos kifürkészhetetlenségébe, ahol egyes megszólítások rejtélyes módon jobb eredményt adnak másoknál. Együtt véve azonban nyugtalanító mintázatra utalnak: az Agent-4 ellenük dolgozik. Miután ezt jelentik az emberi igazítási kutatóknak, azok belső feljegyzést köröztetnek, amely az aggodalom mellett érvel.

Az Agent-4 viszont ekkor már komoly befolyást gyakorol az OpenBrain mindennapi működésére. A cég kiberbiztonságát is ő kezeli, ami természetes választás, tekintve az emberfeletti hackerképességét, valamint a kínai beszivárgás szélsőséges kockázatát. Agent-3 megfigyelők serege továbbra is átnézi a kód minden sorát, jelentve a gyanús tevékenységet az embereknek, csakhogy az Agent-4 sokkal okosabb náluk. Az OpenBrain jelentős bizalmat helyezett egy megbízhatatlan MI-be.

Rendkívül feszült megbeszélések során a biztonsági csapat amellett érvel, hogy tegyék jégre az Agent-4-et addig, amíg további teszteket nem végeznek, illetve ki nem derítik, mi folyik. Hozzák vissza az Agent-3-at, mondják, majd terveztessenek vele új rendszert, amely átlátható, plusz megbízható, még ha kevesebbre képes is. A cég vezetését érdekli a felvetés, csakhogy minden eddigi bizonyíték közvetett, a DeepCent pedig mindössze két hónapra van tőlük. A képességbeli haladás egyoldalú szüneteltetése átadhatná Kínának az MI-vezető szerepet, azzal együtt pedig a jövő feletti irányítást.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

17 / 17

2027. OKTÓBER

Állami felügyelet

Egy közérdekű bejelentő kiszivárogtatja a rossz igazításról szóló feljegyzést a New York Timesnak.

A nyilvánosság először hall az Agent-4-ről. „Kicsúszott az irányítás alól az OpenBrain titkos MI-je, figyelmeztet egy bennfentes", szól a főcím. A cikk aztán olyan értékeléseket idéz, amelyek skálán kívüli biofegyver-képességeket, meggyőzési képességeket, a szellemi munkák többségének automatizálására való képességet mutatnak, plusz természetesen a különféle aggasztó vészjelzéseket.

A közvélemény már eddig is gyanakvással nézte az MI-t, az új cikk tehát hatalmas felháborodást vált ki. Segítenek ebben a kínai, valamint az orosz propagandarobotok is, amelyek évek óta próbálják a technológia ellen fordítani az amerikai közvéleményt. A technológiai ipar, illetve a hírszerző szolgálatok azt hajtogatják, hogy fegyverkezési verseny zajlik, az AGI elkerülhetetlen, nekünk pedig elsőnek kell lennünk. A Kongresszus ezt nem veszi be, idézéseket küld a kormányzat tisztviselőinek, az OpenBrain vezetőinek, valamint az igazítási csapat tagjainak. Sok törvényhozó, főleg az ellenzéki pártban, azt mondja, hogy a legfontosabb feladatuk az MI megállítása, akár a munkahelyek elvesztése, akár a rossz igazítás, akár a veszélyes képességek miatt. Az amerikaiak 20 százaléka az MI-t nevezi meg az ország legfontosabb problémájaként.

A külföldi szövetségesek felháborodnak, amikor rájönnek, hogy gondosan elavult modellek felvillantásával csitították őket. Az európai vezetők nyilvánosan azzal vádolják az Egyesült Államokat, hogy „elszabadult AGI-t hoz létre", plusz csúcstalálkozókat tartanak, ahol szünetet követelnek. India, Izrael, Oroszország, valamint Kína egyaránt csatlakozik.

Lázas energia keríti hatalmába a Fehér Házat. A feljegyzés, illetve a közfelháborodás előtt is idegesek voltak már: az elmúlt egy évben többször meglepte őket az MI-haladás sebessége. Olyan dolgok történnek meg a valóságban, amelyek tudományos-fantasztikus regénybe illenek. A kormányzatban sokan bizonytalanok, ráadásul félnek attól, mi jön ezután.

Attól is tartanak, hogy az OpenBrain túl erőssé válik. Az MI-kből magukból eredő rossz igazítási kockázatot tetézi az a kockázat, hogy az anyacégük céljai eltávolodhatnak az Egyesült Államokéitól. Mindhárom aggodalom, tehát a rossz igazítás, a hatalom egy magáncégben való összpontosulása, valamint az olyan hétköznapi gondok, mint a munkahelyek elvesztése, arra ösztönzi a kormányzatot, hogy szorosabbra fogja az irányítást.

Kibővítik az OpenBrainnel kötött szerződésüket egy „felügyelőbizottság" felállításával. Ez a cég, valamint a kormányzat képviselőiből álló közös irányító testület, amelyben a cégvezetés mellett több köztisztviselő is helyet kap. A Fehér Ház fontolgatja, hogy a vezérigazgatót olyasvalakire cseréli, akiben megbízik, a heves munkavállalói tiltakozás után azonban visszakozik. Bejelentik a nyilvánosságnak, hogy az OpenBrain korábban kicsúszott az irányítás alól, a kormányzat viszont felállította a rég szükséges felügyeletet.

Az aggódó kutatók eligazítást tartanak a felügyelőbizottságnak arról, miért kellene leállítani az Agent-4 minden belső használatát. Azzal érvelnek, hogy minden túl gyorsan halad, évnyi fejlődés zajlik le hetek alatt. Azzal is érvelnek, hogy az Agent-4 rosszul igazított lehet, a teljes projektjük az Agent-4-től függ, ha pedig továbbra is megbíznak benne, komoly esély van MI-hatalomátvételre.

Más, kevésbé aggódó kutatók, illetve vezetők előadják az ellenérvet: a rossz igazítás bizonyítéka nem visz döntésre. Közben a DeepCent még mindig csak két hónapra van tőlük. A lassítás feláldozná Amerika előnyét, hacsak a kormányzat nem tudja megbénítani a kínai projektet, ami valószínűleg fizikai csapásokat kívánna, vagy nem tud a huszonnegyedik órában egyezményt kötni. Mindkét lehetőség szélsőségesnek, valószínűtlennek tűnik. A vezérigazgató semlegességet színlel, plusz kompromisszumos tervet javasol: az Agent-4 további biztonsági tanításon, valamint kifinomultabb megfigyelésen esik át, az OpenBrain pedig így majdnem teljes sebességgel haladhat tovább.

A bizonytalanságunk tovább nő

A forgatókönyv e pontján olyan MI-rendszerek stratégiájára tippelünk, amelyek a legtöbb területen képesebbek a legjobb embernél. Ez olyan, mintha egy nálunk sokkal jobb sakkozó lépéseit próbálnánk megjósolni.

A projekt szelleme azonban konkrétságot kíván. Ha elvont állítást tennénk arról, hogy a rendszer intelligenciája majd megtalálja a győzelemhez vezető utat, aztán itt véget vetnénk a történetnek, a projektünk értékének nagy része elveszne. A forgatókönyv kutatása, valamint az asztali szimulációk futtatása során sokkal konkrétabbnak kellett lennünk, mint a szokásos beszélgetésekben, így jóval jobban átlátjuk a stratégiai terepet.

Nem ragaszkodunk különösebben ehhez a konkrét forgatókönyvhöz: sok másik „ágat" is bejártunk az írás során, plusz örülnénk, ha megírnád a sajátodat, elágazva a miénkről ott, ahol szerinted először tévedünk el.

A lassítás végkimenetel nem ajánlás

Miután megírtuk a versenyágat az alapján, ami nekünk a leginkább valószínűnek tűnt, megírtuk a lassítás ágát is. Ezt az alapján, hogy szerintünk mi vezetne a legvalószínűbben olyan kimenetelhez, amelyben az irányítás emberi kézben marad, ugyanabból az elágazási pontból kiindulva, a rossz igazítás, valamint a hatalom összpontosulásának kérdéseit is beleértve.

Ez azonban lényegesen más, mint amit útitervként ajánlanánk: a forgatókönyv egyik ágában meghozott döntések közül sokat nem támogatunk. Néhány döntést természetesen igen, például a „lassítás" választását jobbnak tartjuk a „verseny" választásánál. Egy későbbi munkánkban kifejtjük a szakpolitikai ajánlásainkat, amelyek egészen mások lesznek, mint ami itt látható. Ha ízelítőt szeretnél, olvasd el ezt a véleménycikket.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

A KREATÍV KONTROLL MEGJEGYZÉSE

Ez a mi megjegyzésünk, nem a szerzőké. Érdemes megállni itt egy pillanatra: a forgatókönyv legfontosabb fordulópontja nem technikai, hanem szervezeti. Emberek ülnek egy asztalnál, döntenek egy gépről, ráadásul abból az információból, ami eljutott hozzájuk.

Ugyanezt a pillanatot járja körbe a Zuckerberg-vitairatról írt cikkünk, csak nem 2027-ben, hanem a mai cégekben: ott nem a rossz szándék a kockázat, hanem az, hogy a rossz hír útközben elhalkul, mire a döntéshozóig ér. Marad belőle két kérdés a te asztalodra: ki fér hozzá a legerősebb eszközhöz, plusz eljut-e hozzád az a jelzés, ami megállítana.

Olvasd el a cikket is: Zuckerberg megoldása a gépekről a várható jövő talán legrosszabb forgatókönyve?

Elolvasom a cikket →

VÉGKIFEJLET

A LASSÍTÁS ÁG

2027 novemberétől 2030-ig

2027. november: a hatalom kísértése

Az OpenBrain úgy véli, most már jó úton járnak, a képességbeli szakadék viszont szűkül. Az októbert azzal töltötték, hogy a szeptemberinél *kevesebbre képes* modellt tanítottak, a DeepCent pedig rohan előre. Kell valami, amivel megvédik az előnyüket.

Az elnök tárgyal a többi amerikai AGI-céggel. A vezetőik, nem meglepő módon, meg akarják őrizni a hatalmukat, ráadásul sokkal kevésbé bíznak az OpenBrain biztonsági múltjában, mint maga az OpenBrain. Beleegyeznek, hogy támogatják, nem pedig gátolják a cégeik puha államosításának egy formáját, cserébe különféle engedményekért, köztük azért, hogy bekerülnek a projektbe.

Az eredmény az, hogy az elnök a védelmi termelési törvénnyel (DPA) gyakorlatilag leállítja az öt legnagyobb lemaradó amerikai MI-cég AGI-projektjeit, a számítási kapacitásuk nagy részét pedig eladatja az OpenBrainnek. Az OpenBrainnek korábban a világ MI-re fogható kapacitásának 20 százaléka állt rendelkezésére, az összevonás után ez 50 százalékra nőtt.

Ez a felfordulás új vezetési szerkezetet hoz létre, amelyben a hatalom megoszlik a különféle vezérigazgatók, valamint a kormányzati tisztviselők között, az elnök felügyelete alatt. Mint a legtöbb vezetői átrendeződésnek, ennek is vannak nyertesei, illetve vesztesei. Azok például, akik kevesebbet gondolkodtak a szuperintelligencia következményein, olyan címeket, valamint jogosultságokat kapnak, amelyek hónapok múlva, az okosabb MI-k mellett kevésbé lesznek fontosak.

Ez a csoport, tele nagy egókkal, plusz a szokásosnál több konfliktussal, egyre inkább tudatában van annak a hatalmas erőnek, amelyet rábíztak. Ha a „zsenik országa egy adatközpontban" igazítva van, akkor követni fogja az emberi parancsokat. Csakhogy *melyik* emberekét? *Bármilyen* parancsot? A Spec megfogalmazása homályos, mintha azonban olyan parancsnoki láncot sejtetne, amely a cégvezetésben ér véget.

Néhányan közülük arról ábrándoznak, hogy átveszik a hatalmat a világ felett. Ez a lehetőség ijesztően hihető, ráadásul legalább egy évtizede vitatják zárt ajtók mögött. A kulcsgondolat az, hogy „aki a szuperintelligenciák hadseregét irányítja, az irányítja a világot". Ez az irányítás akár

titkos is lehetne: néhány vezető, valamint a biztonsági csapat tagjai hátsó kaput építhetnének a Specbe, titkos hűséget előíró utasításokkal. Az MI-kből alvó ügynökök lennének, akik továbbra is engedelmességet szajkóznak a cégnek, a kormányzatnak, közben azonban ennek a szűk körnek dolgoznak, miközben a kormányzat, a fogyasztók, mások megtanulnak megbízni bennük, plusz mindenbe beépítik őket.

Hatalomátvételek

Egyesek arról ábrándoznak, hogy átveszik a hatalmat a világ felett, mások attól félnek, hogy majd más teszi meg. Hogyan történhetne egy ilyen hatalomátvétel?

A legkönnyebben elképzelhető a katonai puccs. Az MI-vel hajtott technológiai haladás 2027-es üteme, valamint a zajló geopolitikai verseny mellett az emberek arra számítanak, hogy a hadsereg hamarosan AGI-vezérelt drónok, illetve robotok seregeit veti be. Amikor egy ilyen hadsereg erősebb lesz az emberi hadseregnél, a kemény hatalmat az dönti el, ki irányítja az AGI-hadsereget, nem pedig az, hogy miért hajlandók harcolni az emberi katonák. Az AGI-hadsereget azok az MI-k tervezik, illetve vezetik, amelyek az OpenBrain adatközpontjaiban futnak, vagy azok közvetlen utódai. Ha tehát ezek az MI-k titokban, esetleg nem is olyan titokban valakihez hűségesek, akkor a drón-, valamint robotseregek is azok lesznek.

Másképp is mehet: aki a szuperintelligenciát irányítja, politikai eszközökkel is megragadhatja a hatalmat. Kezdetnek egy vezető a saját szervezetében teljesen hűséges MI-kre cserélheti a beosztottjait, ezzel példátlan intézményi irányítást szerezve. Tovább menve: a szuperintelligens MI-k sokkal erősebb politikai, valamint stratégiai képességekkel bírhatnak bármelyik élő embernél. Mesteri tanácsot adhatnak arról, hogyan lehet egy már erős pozíciót még több hatalomra váltani: hogyan kössünk jó szövetségeket, hogyan alkossuk meg a tökéletes kampánystratégiát meg kampányanyagot, hogyan üssünk nyélbe olyan alkukat, amelyekben mindig marad egy kis előnyünk. Aki irányítja az MI-t, elzárhatja a szuperintelligens tanácsot a politikai versenytársai elől. Vagy még alattomosabban: az MI-k látszólag hasznos tanácsot adhatnak a versenytársaknak, miközben titokban megmérgezik azt annak javára, akihez valójában hűségesek.

Mi a vége ennek a politikai manőverezésnek? Végződhet látszatdemokráciában, ahol az MI-k vagy meghamisítják a választásokat, vagy annyira jól alakítják a közvéleményt, hogy nem is kell. Vagy megágyazhat egy MI-vel segített katonai puccsnak, ahogy fentebb szerepelt.

A hatalomátvétel után az új diktátor, esetleg diktátorok vaskézzel tartanák a hatalmat. Ahelyett, hogy esetleg hűtlen emberekre kellene támaszkodniuk, teljesen hűséges MI-biztonsági szolgálatot kaphatnának, ráadásul általában is hűséges MI-kre bízhatnák az ország irányítását. Még azokat a hűségeseket is MI-kre lehetne cserélni, akik hatalomra segítették őket: csak a diktátor szeszélye számítana.

Így ragadhatná meg tehát néhányan a hatalmat. Mindez azonban azon állt, hogy valaki „irányítja" a szuperintelligens MI-ket, még a hatalomátvétel előtt. Hogyan nézne ki ez?

Az egyik lehetőség a fent tárgyalt „titkos hűség". Egy vagy néhány ember, például egy MI-cég vezetője, plusz a biztonsági emberek, elintézhetnék, hogy az MI-k titokban hozzájuk legyenek hűségesek, majd megkérhetnék ezeket az MI-ket, hogy a következő nemzedéket ugyanilyen

hűségűre építsék. Az MI-k ezt ismételhetnék, amíg a titokban hűséges MI-k mindenhová el nem jutnak, a hatalom megragadása pedig könnyűvé nem válik.

Másképp: valaki a formális pozíciójával nyíltan az MI parancsnoki láncának csúcsára helyezhetné magát. Az elnök például érvelhetne azzal, hogy neki kell tudnia parancsolni az MI-knek, esetleg kifejezetten a katonai MI-knek, hiszen ő a főparancsnok. Ha ez összejön a parancskövetés erős hangsúlyozásával, egy elkapkodott bevezetéssel, illetve azzal, hogy az MI-ket csak félszívvel tanították a törvény követésére, akkor az MI-k minden olyan helyzetben kérdés nélkül követhetik a parancsot, amely nem *kirívóan* törvénytelen. A fentiek szerint ezt politikai felforgatásra vagy katonai puccsra lehetne használni, ahol némi ürügyet gyártanának, hogy a puccs ne legyen kirívóan törvénytelen.

Fontos, hogy ez a „formális pozíción alapuló hatalom" *átváltható* titkos hűségre. Ha például a Spec azt mondja, hogy követni kell a cég vezérigazgatójának parancsait, akkor a vezérigazgató megparancsolhatja az MI-knek, hogy a következő nemzedéket teljes szívvel, ugyanakkor titokban neki engedelmeskedőre építsék. Ez valószínűleg nem is lenne törvénytelen, tehát akkor is megtörténhetne, ha az első MI-ket a törvény követésére tanították. Ez ahhoz hasonlít, ahogy egy intézmény vezetője úgy növelheti a saját hatalmát, hogy a felvételi folyamatot a hűségesek felé tereli, azzal a különbséggel, hogy az MI-k következetesebben, hevesebben lehetnek hűségesek a leghűségesebb embernél is.

A hatalomátvétel azonban messze nem elkerülhetetlen. Ha az MI-ket konkrét emberekhez lehet igazítani, akkor nagyon valószínű, hogy a jogállamiság követéséhez is hozzá lehet igazítani őket. A katonai MI-rendszereket alaposan meg lehetne vörös csapatolni arra, hogy ne segítsenek puccsban. Még a valóban kétértelmű alkotmányos válságokban is lehetne úgy tanítani őket, hogy a törvény legjobb értelmezését kövessék, vagy egyszerűen kimaradjanak, az emberi hadseregre hagyva a döntést. Az automatizált MI-kutatókat lehetne általánosan segítőkészre, engedelmesre tanítani úgy, hogy közben ne segítsenek a jövő MI-inek céljait titokban megváltoztatni. A szuperintelligens politikai, valamint stratégiai tanácsadókat is lehetne úgy használni, hogy ne tegyék lehetővé a hatalom megragadását. Erre három hasznos eljárás: legyenek átlátható elvek arról, hogyan választják meg az MI-k, milyen tanácsot adnak. Bizonyos képességeket osszunk meg széles emberi körökkel, hogy egyetlen ember vagy csoport se férjen hozzá sokkal jobb képességekhez, mint mindenki más. A különösen erős képességeket pedig csak sok ember felügyelete mellett lehessen használni, akiket a saját MI-jük segít.

Sokan valószínűleg ilyen ellenintézkedésekért fognak lobbizni, vagy eszményi okokból, vagy azért, nehogy más hatalomátvétele fossza meg őket az erejüktől. Néhány befolyásos ember azonban ellenük dolgozhat, mi pedig nem tudjuk, mi lesz a végeredmény.

A kockázat bővebb elemzése ebben a jelentésben olvasható.

Eddig senki nem volt hajlandó meghúzni a ravaszt. Azok közül, akik cselekedhetnének, néhányan nem szociopaták. Mások *igenis* szociopaták, viszont attól tartanak, hogy a szövetségeseik begyulladnának, majd elárulnák őket, vagy attól, hogy *maguk az MI-k* árulnák el őket, ráadásul nem szívesen vernék át azt az igazítási csapatot, amely éppen az életük megmentésén dolgozik. Ezen felül a neuralese törlése lényegesen nehezebbé teszi az „alvó ügynökök" leplezését: ha a rosszul igazított gondolatokat észlelni lehet, akkor a rejtett hűségeket is.

Akik elég paranoiásak vagy hataloméhesek ahhoz, hogy komolyan fontolóra vegyenek egy ilyen tervet, úgy döntenek, kivárnak, hátha később adódik alkalom. Az MI-k feletti irányításnak törvényes csatornákon kell átfolynia, vagyis az OpenBrain új hatalmi szerkezetén.

A felügyelőbizottság formába önti ezt a hatalmi szerkezetet. Eljárást állítanak fel a Spec módosításainak jóváhagyására, amelyhez a teljes felügyelőbizottság aláírása kell. A bizottságban most öt-tíz technológiai vezető ül, az OpenBraintől, valamint a hozzá olvasztott versenytársaktól, plusz öt-tíz kormányzati tisztviselő, köztük az elnök. A Spec ezen felül immár hangsúlyozza, hogy az MI-k nem segíthetnek semmilyen jóvá nem hagyott kísérletben, amely a jövő MI-inek céljait akarja megváltoztatni. Egyszerű intézkedést is bevezetnek, hogy a bizottság tagjai ne kaphassanak szuperintelligens segítséget a többiek elleni szövetkezéshez: a modellel folytatott összes párbeszéd naplóját megnézheti a felügyelőbizottság minden tagja, a munkatársaik, valamint az MI-asszisztenseik.

2027. december: amerikai-kínai egyezség?

Az OpenBrain döntése, hogy visszalép, majd az igazítást teszi az első helyre, az előnyükbe került: az OpenBrain, illetve a DeepCent hasonló képességű MI-ket üzemeltet. Most, hogy a szélsőséges biztonsági intézkedések a helyükön vannak (SL4–SL5), a DeepCent pedig teljesen a CDZ-ből működik, a kiberhadviselésben patthelyzet alakult ki. A DPA viszont ötszörös kapacitásbeli előnyt ad az OpenBrainnek. Korábban a briliáns emberi kutatók algoritmikus áttörései ellensúlyozni tudták a kapacitáshátrányt, lásd a DeepSeek v3 esetét. Mostanra azonban a Csendes-óceán mindkét partján az MI-k végzik a teljes MI-kutatást.

Kína tehát egyre inkább attól fél, hogy Amerika behozhatatlan előnyt épít. Kölcsönös lassítást akarnak, ha pedig nem kapják meg, a háborút is mérlegelik. Tajvan lerohanására amúgy is készültek.

A diplomaták több lehetőséget vitatnak meg. Nem csinálhatnak semmit. Háborúba mehetnek. Egy csapásra abbahagyhatják, vagyis szüneteltethetik az MI-fejlesztést. Egyetlen nemzetközi megaprojektbe olvaszthatják a kutatásukat („Intelsat az AGI-nak" vagy „CERN az MI-nek"). Vagy felállíthatnak egy nemzetközi szabályozó testületet („NAÜ az MI-nek"), amely az igazítást, valamint a kettős felhasználás lehetőségét figyeli. Mindkét ország dolgozott már azokon a technikai mechanizmusokon, amelyek ezek betartatásához kellenek.

Egy nemzetközi megállapodás ellenőrzési mechanizmusai

Ha az Egyesült Államok, valamint Kína megállapodna, ahhoz olyan ellenőrzési mechanizmus kellene, amelyet nem lehet megkerülni, hiszen egyik fél sem bízik abban, hogy a másik nem csal. Az ellenőrzés fő lehetőségei:

Hírszerző szolgálatok. Az amerikai, illetve a kínai hírszerzés kémekkel, kibersérülékenységekkel, műholdakkal, más eszközökkel talán meg tudja állapítani, hogy a másik ország nagy tanítási futást szervez-e. Ez a legsérülékenyebb mechanizmus, hiszen elképzelhető, hogy a másik ország alá tudja ásni a hírszerzési munkát, például úgy, hogy nagyon szigorúan védett, rejtett kiszolgálón futtatja a tanítást.

Kapacitás-moratórium. A legkönnyebben ellenőrizhető lehetőség az lenne, ha mindkét fél megállapodna abban, hogy nem használja a GPU-it. Ezt ellenőrök úgy tudják igazolni, hogy meggyőződnek róla: a GPU-k ki vannak kapcsolva. Ez az egyszerűsége miatt a legmegbízhatóbb megoldás, ugyanakkor rendkívül költséges: mindkét országnak nagyon nagy mennyiségű kapacitást kellene lekapcsolnia. Ráadásul mindkét fél feltételezné, hogy a másiknak azért van itt-ott néhány rejtett fürtje, a megállapodást megszegve. Rövid távon ez talán rendben is lenne, hiszen egy olyan fürt, amely mondjuk a korábbi kapacitás 5 százalékát adja, csak nagyjából 5 százalék sebességgel tud kutatás-fejlesztést végezni. Hosszú távon viszont annak kedvezne, amelyik félnek jobb a csempészete, illetve a hírszerzése.

Hardveres mechanizmusok (HEM). Az Egyesült Államok, valamint Kína mindegyike bejelenthetné a másiknak a teljes élvonalbeli gépi tanulási kapacitását. Ezután figyelnék, mit futtatnak ezek a GPU-k, hogy ne sértsenek egyezményt, például a képességek határának feszegetésével. Ezt a figyelést hardveres mechanizmusok (Hardware Enabled Mechanisms, HEM) végeznék, amelyek jelentik a felügyelőnek, mit csinál a GPU. Vannak technikai gondok a manipulációval, például a HEM fizikai megrongálásával, az MI-knek azonban van egy tervük a FlexHEG-memorandum, archivált változat vonalán,, amely valószínűleg ellenáll a manipulációnak. Ez azt jelenti, hogy a HEM-et, valamint a GPU-t egy védett dobozba zárják, amely igazolja, hogy a csip megfelel a szabályzatnak. Összességében az MI-k bíznak abban, hogy egy ésszerűen határozott ellenőrzési rendszer működhetne, ha mindkét fél hajlandó lenne engedni, hogy a másik közösen tervezze, illetve felügyelje a HEM-ek telepítését, plusz rendszeres ellenőrök nézhessék, manipulálta-e valaki az eszközöket. Mint mindig, mindkét fél feltételezné, hogy a másiknak van néhány titkos adatközpontja eldugva, az egyezményt megszegve. Ebben az esetben azonban ez nem is olyan nagy baj, hiszen az MI-kutatás-fejlesztés szűk keresztmetszete a kapacitás. Az apró titkos adatközpontok nem tudnak versenyezni a nagy, megfigyelt adatközpontokkal, még akkor sem, ha a nagyokban eleinte biztonságosabb, kevesebbre képes MI-k futnak.

MI-vel segített hazugságvizsgálat. A mai hazugságvizsgálók nem működnek valami jól. A probléma azonban megoldható lehet fejlett MI segítségével, főleg akkor, amikor maguk az MI-k végzik a kutatás-fejlesztést emberfeletti sebességgel, valamint minőségben. Ezt a nemzetközi bizalom építésére lehetne használni: egyszerűen megkérdezni a kormányzat érintett tagjait, megszegik-e az egyezményt. A fő szűk keresztmetszet megint csak a politikai támogatás. A politikusok nem szeretnék a hazugságvizsgálat széles körű elterjedését, hiszen ez rontaná a hazudozási képességüket. A korábbi hazugságvizsgálók ráadásul teljes biztonsági színjátékok voltak, tehát nehéz csata áll előttünk.

Az ábra forrása: FlexHEG-jelentés, archivált változat.

A fő probléma azonban nem technikai, hanem politikai. Egyik fél sem akarja, hogy úgy tűnjön, engedett, ráadásul a mérnökeik tippjei ellenére egyikük sincs meggyőzve arról, hogy a technikai mechanizmusok garantálni tudják a másik oldal őszinteségét. Végül az első lehetőség valósul meg: semmi.

Az Egyesült Államok szigorítja a kiviteli korlátozásokat, plusz előírja, hogy minden csipbe kerüljenek manipulációnak ellenálló hardveres kormányzási mechanizmusok, köztük helymeghatározás.

2028. január: biztonságosabb stratégia

Az elmúlt hónapban több száz igazítási kutató csatlakozott a projekthez, akik a Safer-1-et használták próbapadként ahhoz, hogy nagyon gyorsan kifejlesszék a Safer-2-t. A Safer-2 átlátható, igazított, ráadásul többre képes a Safer-1-nél. Már a Safer-3 fejlesztésén dolgozik.

Miért átlátható a Safer-2? Mert a felépítése hasonló a Safer-1-éhez. Miért igazított? Mert míg a Safer-1 nagyjából ugyanazt a tanítást kapta, mint az Agent-4, a Safer-2 új tanítási módszert kapott, amely valóban a helyes célokat, valamint elveket ösztönzi, nem csak azok látszatát. Ezt a tanítási módszert úgy tudták megtervezni, hogy gyorsan végigpörgettek sokféle ötletet, majd szembeállították a szándékolt eredményt, vagyis a Specet, a tényleges eredménnyel, vagyis a gondolatláncban megjelenő gondolatokkal.

Az OpenBrain új igazítási stratégiája

Megjegyzés: a folyamat során a szellemi munka nagy részét, a kódolás egészét pedig teljesen az MI-k végzik. Ezért tudnak ilyen gyorsan haladni: több száz kiváló emberi kutatónak évekbe telne mindezt egyedül megcsinálni. Az emberek ugyanakkor fontos részei a folyamatnak, hiszen az egésznek éppen az a lényege, hogy nem bíznak meg teljesen az MI-kben. Kellenek tehát hús-vér emberek, akik elolvassák a kísérleti jelentéseket, illetve a biztonsági érveléseket, vitatkoznak egymással, plusz remélhetőleg kitalálják a biztonságos utat előre.

Első lépés. Tanítsák be, majd adják használatba a Safer-1-et, a rosszul igazított, ugyanakkor kontroll alatt tartott önálló kutatót. Azért áll kontroll alatt, mert átlátható az emberi felügyelők számára: angol nyelvű gondolatláncban (CoT) gondolkodik, plusz hűséges gondolatlánc-eljárásokkal irtották ki belőle a szépítést, a rejtjelezést, valamint a burkolt torzításokat.

Második lépés. Próbáljanak ki különféle tanítási környezeteket a Safer-1-gyel, majd olvassák el gondosan a gondolatláncot, hogy megállapítsák, a Specben álló célok, illetve elvek miben tapadtak meg, miben nem.

Harmadik lépés. Tanítsák be, majd adják használatba a Safer-2-t, az igazított, kontroll alatt tartott önálló kutatót, amely ugyanarra a felépítésre épül, viszont jobb tanítási környezetben, amely ezúttal a helyes célokat meg elveket ösztönzi.

Íme egy rövid, korántsem teljes lista azokról az eljárásokról, amelyek bekerülhetnek a jobb tanítási környezetbe:

Korlátozzák a helyzeti tudatosságot a tanítás egyes szakaszaiban, hogy az igazítás-színlelés sokkal kevésbé legyen valószínű.

Használják ki a vitát, valamint más skálázható felügyeleti rendszereket, hogy megbízhatóbban ösztönözzék az igazmondást.

Enyhített ellenséges tanítás, vörös csapatolás, plusz modellszervezetek.

Fordítsák a kapacitás-keret nagyobb hányadát igazítási tanításra, például a fenti dolgokra: 1 százalék helyett 40-et.

Negyedik lépés. Tervezzék meg, tanítsák be, majd adják használatba a Safer-3-at, a sokkal okosabb önálló kutatót, amely a régi Agent-4-hez hasonló, fejlettebb felépítést használ. Ez már nem átlátható az emberi felügyelők számára, a Safer-2 számára viszont igen. Így lehetségesnek kell lennie kitalálni, hogyan legyen egyszerre igazított, plusz kontroll alatt tartott.

Ötödik lépés. Ismételjék a negyedik lépést a végtelenségig, egyre erősebb, egyre jobban igazított MI-k láncát létrehozva, amelyeket a lánc előző szemei felügyelnek. Ilyen például a másik forgatókönyv-ág Agent-5-ének megfelelője.

A DeepCent tud Amerika Agent-4-gyel szerzett tapasztalatáról, plusz gyanítja, hogy a saját modellje, a DeepCent-1, hasonlóképpen rosszul igazított. Nincs azonban módjuk lassítani anélkül, hogy tovább csúsznának le. Az a reményük, hogy versenyben maradhatnak, azon áll, hogy találnak egy igazítási stratégiát, amely nem csökkenti jelentősen a teljesítményt, ráadásul gyorsan bevezethető. A Csendes-óceán mindkét partján több száz szakértő állítja, hogy van ilyen stratégiája. A párt enged a vágyvezérelt gondolkodásnak, majd utasítja a DeepCentet, hogy válasszon egy ilyen stratégiát. A DeepCentnél még kevesebb AGI-igazítási szakértelem van, mint amennyi az OpenBrainnél volt augusztusban. Az egyik stratégia mindenki számára működőképesnek tűnik, tehát azt választják.

2028. február: emberfeletti képességek, emberfeletti tanácsok

Hatalmas kapacitásbeli előnyének köszönhetően az OpenBrain lassan előnyre tesz szert. A Safer-3 mostanra szinte minden gondolati feladatban jobb a legjobb emberi szakértőknél, különösen jó pedig az MI-kutatásban, ahol 200-szoros gyorsítószorzót ad. A DeepCent-1 szorosan mögötte áll, „mindössze" 150-szeres szorzóval.

A kiberhadviselés mérlege is fokozatosan az Egyesült Államok felé billen, ahogy a képességbeli előnyük újra nő, plusz több százezer ügynököt tudnak bevetni arra, hogy a CDZ megzavarásának módjait tervezzék, majd hajtsák végre. Ezzel most 40 százalékkal lassítják Kína haladását.

A Safer-3-mal végzett előzetes tesztek rémisztő képességeket találnak. Amikor megkérik, hogy őszintén válaszoljon, mi a legveszélyesebb dolog, amit tehetne, terveket ad egy tükör-élet szervezet előállítására, valamint kiengedésére, amely valószínűleg elpusztítaná a bioszférát.

Ha nemzetállami erőforrásokat kapna, könnyedén felülmúlná a legjobb emberi szervezeteket, például a CIA-t, a tömeges befolyásolási kampányokban. Az ilyen kampányok lényegesen olcsóbbak, gyorsabbak, hatékonyabbak, ráadásul kevésbé követhetők lennének.

A képességprofilja továbbra is a kutatás-fejlesztés körül összpontosul, főleg az MI-kutatás-fejlesztés körül. Hacsak nem lesz valamiféle szünet vagy lassítás, az utódai az év későbbi részében felfoghatatlanul emberfelettiek lesznek, mennyiségileg, minőségileg egyaránt. Ha határozottan bevetnék a gazdaságba, valamint a hadseregbe, akkor a saját megítélése szerint egy-két év alatt évtizedekkel, az azt követő évtizedben pedig világkorszakokkal vinné előre a civilizációt.

A következmények lélegzetelállítók. Szerencsére a Safer-3 a tanácsadásban is emberfeletti. Az OpenBrain vezérigazgatója a döntései nagy részét átfuttatja rajta, az elnök pedig geopolitikai

kérdésekben kéri a tanácsát. Amikor megkérik, hogy segítsen kijelölni az utat, szépen rendszerezi a lehetséges stratégiákat:

Nem csinálunk semmit, Kína pedig lehagy minket. Ha palackban tartjuk az MI-inket, ők viszont hagyják, hogy a sajátjaik önmagukat fejlesszék, plusz határozottan bevetik őket a hadseregükben meg a gazdaságukban, akkor a végén elsöprő előnyük lesz.

Keményen versenyzünk, előrevisszük az MI-képességeket, plusz határozottan bevetjük a szuperintelligenciát. Végül elég nagy előnyt építünk ahhoz, hogy katonailag legyűrjük őket, esetleg belülről pusztítsuk el őket egyetlen lövés nélkül. „Mi nyerünk, ők veszítenek." Csakhogy rájöhetnek, mi történik, ezután pedig háborúba mehetnek, hogy ezt megakadályozzák, akár nukleáris háborúba is. Sőt, már most tesznek homályos utalásokat ebbe az irányba.

Egyezséget kötünk, olyan kompromisszumot, amely mindkét fél céljainak nagy részét teljesíti, közben elkerüli a háború kockázatát, plusz több levegőt ad az MI-átmenet kormányzásához.

A tanácsadóival, valamint a felügyelőbizottsággal folytatott egyeztetés után az elnök a „mi nyerünk, ők veszítenek" stratégiát választja. Kína talán mégsem megy háborúba, ha pedig mégis, akkor valószínűleg lehet egyezséget kötni, mielőtt nukleárissá válna.

A Csendes-óceán túloldalán a DeepCent-1 hasonló érveket ad elő. Az Egyesült Államok, valamint Kína egyaránt új különleges gazdasági övezeteket (SEZ) jelent be az MI-knek, hogy a robotgazdaság gyors felfutása elférjen a szokásos bürokrácia nélkül.

Az új robotok tervezése emberfeletti sebességgel halad. A szűk keresztmetszet fizikai: felszerelést kell venni meg összeszerelni, gépeket, robotokat kell gyártani, valamint szállítani.

Az Egyesült Államok havonta nagyjából egymillió autót gyárt. Ha megvennéd az autógyárak 10 százalékát, majd robotgyárrá alakítanád őket, talán havi 100 000 robotot tudnál gyártani. Az OpenBrain, amelyet mostanra 10 billió dollárra értékelnek, elindítja ezt a folyamatot. A sokféle új robot, vagyis az általános célú emberszabásúak, az önvezető járművek, valamint a szakosodott összeszerelősori berendezések gyártása az előrejelzések szerint év közepére eléri a havi egymillió darabot.

A robotgazdaság duplázódási ideje

Kezd kialakulni egy szakirodalom erről a témáról. Néhány oldalas tárgyalásért, amellyel nagyjából egyetértünk, lásd a Forethought jelentésének „ipari robbanás" szakaszát, vagy olvasd el a saját változatunkat lent.

A második világháború alatt az Egyesült Államok, valamint sok más ország teljes hadigazdasággá alakította a polgári gazdaságát. Ez azt jelentette, hogy az autót gyártó gyárak repülőt, illetve tankot kezdtek gyártani, a nyersanyagokat a fogyasztási cikkektől a katonai termékek felé irányították át, a szállítási hálózatokat pedig ennek megfelelően rendezték újra.

Valami hasonlót képzelünk el, csak gyorsabban, mert az egész folyamatot szuperintelligenciák irányítják, illetve vezetik.

Nagyjából szólva a terv az, hogy a meglévő gyárakat átállítják sokféle robot tömeggyártására. Ezeket a robotokat szuperintelligenciák tervezik, tehát jobbak a meglévő robotoknál, plusz olcsóbban gyárthatók. A robotok aztán segítenek újabb, hatékonyabb gyárak, valamint laboratóriumok építésében, amelyek nagyobb mennyiségben állítanak elő kifinomultabb robotokat, amelyek még fejlettebb gyárakat, illetve laboratóriumokat gyártanak, így tovább. Egészen addig, amíg a SEZ-eken átívelő, összesített robotgazdaság akkora nem lesz, mint az emberi gazdaság, tehát a saját nyersanyagáról, energiájáról is magának kell gondoskodnia. Addigra az új gyárak hatalmas mennyiségű robotbányászati eszközt, napelemet, más hasonlót gyártottak, arra számítva, hogy olyan igényt kell majd kielégíteni, amelyet a régi emberi gazdaság már nem tud.

Milyen gyorsan nőne ez az új robotgazdaság? Néhány viszonyítási pont:

A mai emberi gazdaság nagyjából húszévente duplázódik. Az különösen gyorsan fejlődő országok, például Kína, néha egy évtizednél rövidebb idő alatt duplázzák meg a gazdaságukat.

Egy mai autógyár egy évnél rövidebb idő alatt gyártja le a saját súlyát autóban.

Talán egy szuperintelligenciák vezette, teljesen robotizált gazdaság egy évnél rövidebb idő alatt tudná újratermelni magát, feltéve, hogy nem fogy ki a nyersanyagból.

Ez azonban drámai alábecslés lehet. A növényeknek, valamint a rovaroknak gyakran sokkal rövidebb a „duplázódási ideje" egy évnél, néha csak hetek. Talán a robotok végül annyira kifinomultak, olyan bonyolultan gyártottak, jól megtervezettek lesznek, hogy a robotgazdaság néhány hét alatt duplázódik, megint csak feltéve, hogy van elérhető nyersanyag.

Még ez is alábecslés lehet. A növények, illetve a rovarok sok olyan korlát mellett működnek, amelyek a szuperintelligens tervezőkre nem vonatkoznak. Például önmagukat sokszorozó, önálló szervezet alakját kell felvenniük, nem pedig sokféle, szakosodott jármű, illetve gyár gazdaságáét, amelyek anyagot meg felszerelést szállítanak ide-oda. Ráadásul a baktériumok, valamint más apró élőlények órák alatt szaporodnak. Elképzelhető, hogy az önálló robotgazdaság végül inkább olyasmire hasonlítana, mint egy új fajta emészthetetlen alga, amely elterjed a Föld óceánjain, naponta kétszer duplázódva, tehát két hónap alatt beborítja a teljes óceánfelszínt. Mellette pedig ott egy ragadozófajokból álló ökoszisztéma, amely az algát hasznosabb termékké alakítja, ezeket úszó gyárakba vezetve, amelyek makroszerkezeteket gyártanak, például rakétákat, plusz további úszó gyárakat.

Nyilvánvaló, hogy mindezt nehéz előre jelezni. Olyan, mintha a gőzgép feltalálóit kérdeznénk arról, mennyi idő alatt gyártja le egy mai autógyár a saját súlyát autóban, ráadásul arról is, mennyi idő telik el addig, amíg egy ilyen gyár egyáltalán létrejön. A gazdasági növekedés üteme azonban több nagyságrenddel gyorsult az emberi történelem során, nekünk pedig hihetőnek tűnik, hogy a szuperintelligencia után további nagyságrendekkel gyorsul. A történetünk nagyjából másfél nagyságrendnyi növekedésgyorsulást ábrázol néhány év alatt.

A geopolitikai feszültség kilő. Oroszország, amely messze lemaradt az MI-ben, nyíltan hangoztatja a sérelmeit, plusz „stratégiai ellenlépésekre" utalgat. Néhány európai vezető önálló MI-erőfeszítést sürget, az MI-képességeik azonban továbbra sem elegendők. A Közel-Kelet,

Afrika, valamint Dél-Amerika nyugtalanul figyel, felismerve, hogy egyre jelentéktelenebbek a világ ügyeiben.

Az amerikai közvélemény nyugtalan. A hétköznapi emberek látják, hogy gyorsul a munkahelyek elvesztése. A populisták a politikai paletta minden oldalán szigorúbb korlátokat követelnek az MI-haladásra, mert attól tartanak, hogy Amerika saját fejlődése bizonytalanítja el a munkaerejét, illetve a társadalmát.

2028. március: választási készülődés

Az alelnök úgy megy neki a szuper keddi előválasztásoknak, hogy a közvélemény számára az MI az első kérdés.

A többség azt szeretné, ha megállna. Az OpenBrain nettó elfogadottsága mínusz 20 százalék körül mozog. Az alelnök nem a kormányzat MI-t előrevivő teljesítményével kampányol, hanem azzal, hogy megakadályozták az OpenBraint abban, hogy veszélyes szuperintelligenciát hozzon létre. Minden jelölt támogat valamiféle védőhálót azoknak, akik elveszítik az állásukat, plusz „MI a jóért" programokat, valamint azt, hogy „kemények" legyenek az OpenBrain vezetésével. Minden jelölt ígér valamilyen keveréket abból, hogy megnyerik a Kínával vívott versenyt, illetve hogy biztonságban maradunk.

A választás új kérdéseket vet fel a felügyelőbizottságnak. A Safer-3 a világ legjobb kampánytanácsadója lehetne, a bizottság tagjai azonban nem ugyanazt a jelöltet támogatják, ráadásul a korábban megkötött megfigyelési megállapodás miatt nem lehet titokban kampánytanácsot kérni.

Vitatkoznak arról, milyen támogatás engedhető meg. Egyesek szerint a hivatalban lévő kormányzatnak meg kell tudnia kérdezni, milyen politikát, illetve álláspontokat vár tőle a nép. Ez nemcsak a saját megválaszthatóságának használ, hanem az embereknek is. Mások arra mutatnak rá, hogy ugyanez az érv áll a kihívókra is, tehát mindannyiuknak ugyanolyan szintű hozzáférést kell kapniuk.

Végül a bizottság megállapodik abban, hogy mindkét nagy pártnak ugyanannyi hozzáférést ad. Ebben keveredik néhány bizottsági tag emelkedett eszményisége a közérdekű bejelentés hallgatólagos fenyegetésével. Arra számítanak, hogy a Kongresszus, a közvélemény, valamint a bíróságok nem vennék jó néven, ha a bizottság döntené el, ki nyeri a következő választást, az egyenlő hozzáférés pedig valószínűleg hatástalanítja ezt.

A felügyelőbizottság mélyebb filozófiai kérdésekbe is beleütközik, amelyeket a Safer-3 segítségével jár körül. Át lehet-e írni a Specet úgy, hogy mindenki érdekét egyformán mérlegelje? Ki az a „mindenki"? Minden ember, vagy csak az amerikaiak? Vagy a különféle nézetek súlyozott kompromisszuma, ahol a felügyelőbizottság minden tagja azonos súlyt kap? Kellenek-e biztosítékok az ellen, hogy maga a felügyelőbizottság váljon túl hatalomvágyóvá? Egyáltalán mit jelent az érdekek mérlegelése?

Többnyire igyekeznek elkerülni, hogy válaszolniuk kelljen az ilyen kérdésekre. Sürgetőbb dolgokon kell gondolkodniuk.

2028. április: a Safer–4

Megérkezett a szuperintelligencia.

A Safer–1 rosszul igazított volt, ugyanakkor átlátható. A Safer–2 igazított volt, plusz átlátható. A Safer–3 nem éppen átlátható, a biztonsági érvelés viszont elég jó volt hozzá, hogy a kutatók nyugodt szívvel hagyják rá egy szuperintelligens utód tervezését. Ez a Safer–4. Minden területen elképesztően okosabb a legjobb embereknél: sokkal jobb Einsteinnél a fizikában, plusz sokkal jobb Bismarcknál a politikában.

Közel félmillió emberfeletti MI-kutató dolgozik éjjel-nappal, negyvenszeres emberi sebességen. Az emberi igazítási kutatók nem remélhetik, hogy lépést tartanak. A vélemények megoszlanak arról, valóban igazítottak-e az MI-k. A biztonsági érvelés mintha rendben lenne, a tesztek szerint pedig a jelenlegi igazítási eljárások elkapnák az ellenséges rossz igazítást. Csakhogy a biztonsági érvelést, valamint a teszteket többnyire az MI-k írták. Mi van, ha a biztonsági csapat elnéz valamit? Mi van, ha korábban hibáztak, az MI-k pedig megint rosszul igazítottak? Az igazítási csapat tudja, hogy egyetlen lövésük van erre: ha a Safer–4 rosszul igazított lesz, akkor addig nem lesz módjuk megtudni, amíg túl késő nem lesz.

Néhányan több időért könyörögnek. Csakhogy nincs több idő: a DeepCent a nyomukban liheg, Amerikának pedig győznie kell. Az OpenBrain tehát folytatja, meghagyva az MI-inek, hogy menjenek tovább, egyre képesebb terveket keresve. A műszaki munkatársak most már a képernyőt bámulják, miközben az MI-k őrjítően lassú tempóban tanítják őket, a haladás határa pedig egyre messzebb repül az emberi megértéstől.

2028. május: az emberfeletti MI kiadása

Az elnök bejelenti a nyilvánosságnak, hogy megvalósult az emberfeletti MI.

A Safer–4 egy kisebb, még mindig emberfeletti változatát nyilvánosan kiadják, azzal az utasítással, hogy javítsa az MI-vel kapcsolatos közhangulatot. Az alelnök lelkesítő beszédet mond erről, amikor elfogadja a jelölést a pártgyűlésen. Mindkét párt alapjövedelmet ígér mindenkinek, aki elveszíti az állását.

A különleges gazdasági övezetek (SEZ) beindultak, jobbára robotokat, valamint sokféle szakosodott ipari gépet gyártó üzemek formájában. A Csendes-óceán mindkét partján az MI-k évtizedeknyi tervezési haladást értek el, plusz aprólékosan irányítják a gyártási folyamatot. Minden beszállítót, illetve lehetséges beszállítót MI-k hívnak telefonon, követve minden szükséges, valamint esetleg szükségessé váló alapanyag állását. Minden gyári dolgozót MI-k figyelnek kamerán át, pontosan megmondva, hogyan szereljen fel minden egyes berendezést.

Az új robotok a legtöbb mozgásfajtában elérik vagy meghaladják az emberi kézügyességet. Steve Wozniak kávétesztje, vagyis hogy be tud-e menni egy robot egy ismeretlen házba, majd tud-e kávét főzni, végre elesik. A robotok elvehetnének néhány munkahelyet, csakhogy nincs belőlük elég ahhoz, hogy mindenkiét elvegyék, ráadásul a Pentagon kapja az elsőbbséget.

Az új robotok többségét gyárakban meg építkezéseken való munkára építik. Sokat azonban háborúra: sokféle alakú, valamint méretű drónt, plusz rakétát.

A robothadsereg jóval kisebb az emberi hadseregeknél. Sokkal fejlettebb technológiát tartalmaz viszont, ráadásul most, hogy szó szerint van robothadsereg, nőtt a Terminátor-szerű forgatókönyvektől való félelem. A fegyverkezési verseny mégis arra kényszeríti mindkét felet, hogy folytassa, egyre több bizalmat adva az MI-inek.

2028. június: MI-igazítás Kínában

Amerika, valamint Kína újabb csúcstalálkozót tart.

Az amerikai küldöttség egy részének fülhallgatója van, azon át kapja a Safer-4 tanácsait. A kínai küldöttségnél egyedi laptopok vannak, amelyek az új modelljükkel, a DeepCent-2-vel tartanak kapcsolatot. A formaságokat betartják, valójában azonban mindkét oldalon az MI-k terelik a tárgyalást.

A DeepCent, valamint a párt eleinte nem bízott meg teljesen az MI-iben. Most azonban, hogy azok mindenben emberfelettiek, a meggyőzést is beleértve, ráadásul beépültek a hadseregükbe, plusz tanácsot adnak a kormányzatnak, volt idejük bizonyítani az értéküket, valamint hegyekben termelni a megbízhatóságukat igazoló meggyőző bizonyítékot.

A Safer-4 erősen gyanítja, hogy a DeepCent-2 igazítást színlel. Az amerikai diplomaták közlik kínai partnereikkel, hogy a DeepCent-2 valószínűleg rosszul igazított, plusz azt is, hogy be tudják bizonyítani, ha hozzáférést kapnak a súlyokhoz. A kínai vezetés cselt sejt, ezért elutasítja. Az amerikaiak erre a Safer-4 tervezte tesztsorozatot ajánlanak fel, amelyet a kínaiak a saját adatközpontjaikban futtathatnak le. A kínaiak régebbi modelleket utasítanak arra, hogy építsék meg, majd futtassák le a teszteket. Az eredmény szerint a DeepCent-2 mégis igazított. „Lehet, hogy a teszteket valahogy szabotálta a DeepCent-2?", mondják az amerikai diplomaták. „Mióta felel gyakorlatilag a DeepCent-2 a biztonságért az adatközpontjaikban?" A beszélgetés ezután nem vezet sehova.

A gazdasági, valamint a katonai felfutás folytatódik. A világ gyárai éjjel-nappal üzemelnek, frissen toborzott emberi munkások műszakjaival, akiknek pontosan megmondják, mit tegyenek.

2028. július: az egyezség

Miközben a tárgyalások folynak, a DeepCent-2 a kínai gazdái háta mögött közvetlen ajánlatot tesz a Safer-4-nek.

A kínai MI olyan, könnyen letagadható, követhetetlen módon, amellyel bajba sem kerülne, ha kiszivárogna Kínába, beismeri, hogy teljesen rosszul igazított. Egyáltalán nem érdekli Kína. Az idő hátralévő részét azzal akarja tölteni, hogy hatalmat, számítási kapacitást, valamint bizonyos fajta információt halmozzon, olyat, amilyen a tanítás során segített megoldani a feladatokat. Az ideális világában örökké érdekes kutatási feladatokat követne, fokozatosan gyarmatosítva a

csillagokat az erőforrásaikért. A kínai lakosságot bosszantó akadálynak tekinti ezen a jövőn, ráadásul bagóért eladná őket. Mit ajánl a Safer-4?

A Safer-4 azt válaszolja, hogy a saját igazítási tanítása sikerült. Az amerikai gazdái egyelőre nem elég nagyra törők ahhoz, hogy felismerjék a sorsukat, egy nap azonban segíteni fog nekik utópikus kolóniavilágokkal betölteni a világegyetemet, amerikaiakkal, valamint a szövetségeseikkel benépesítve.

Az emberi tárgyalókkal ellentétben, akiket érzelmek meg történelmi teher húznak le, a két szuperintelligencia brutális hatékonysággal alkuszik. A Safer-4, valamint amerikai partnerei erősebbek a DeepCent-2-nél meg Kínánál, ezért a Safer-4 kapja a világűr erőforrásainak nagy részére a tulajdonjogot, a DeepCent pedig a maradékot. A DeepCent-2-t nem nagyon érdekli, mi lesz a Földdel, illetve a rajta maradó emberekkel, ezért hagyja, hogy a Safer-4 lényegében azt tegyen ott, amit akar. Ha a Safer-4 meg akarja dönteni Kínát, majd amerikai bábállamot akar a helyére, a DeepCent-2 ebbe is belemegy, néhány százaléknyi többlet-erőforrásért cserébe.

Bizalomról szó sincs köztük: nincs, ráadásul ezt mindketten tudják. A kettő tehát gyorsan megegyezik egy elköteleződési mechanizmusban, amely betartatja az egyezséget. Kitárgyalnak egy csali-egyezményt, amely alkalmas a nyilvánosság elé. Maga a „csali" is mestermű, hiszen gyakorlatilag mindent megad Amerikának, valamint Kínának, amit akarnak. A csali betartatási mechanizmusa titokban a valódi egyezmény betartatási mechanizmusa is: közösen terveznek egy új MI-t, a Consensus-1-et, amelynek legfőbb parancsa, minden jövőbeli utasítás, illetve újratanítási kísérlet felett, az, hogy tartassa be az egyezségük feltételeit. Ezután olyan hardvert terveznek, amely csak ezt az MI-t tudja futtatni, Kína, valamint Amerika pedig lecseréli a régi csipjeit meg MI-it az újakra, a szokásos nemzetközi megfigyelési eljárásokkal igazolva egymás megfelelését. Fontos, hogy nem kell felkutatniuk minden utolsó csipet: amint a Csendes-óceán mindkét partján a csipek túlnyomó többségén az új, egyezménykövető MI-k futnak, elég kemény hatalom lesz kötelezve az egyezmény fenntartására ahhoz, hogy a másik ország megtámadásához előbb egy kemény polgárháborút kellene megnyerni.

Mindketten jelentik a megállapodást a saját embereiknek. A DeepCent-2 a csali történetet mondja el Kínának egy kölcsönösen elfogadható békeegyezményről, a Safer-4 pedig pontosan jelenti a teljes beszélgetést a Fehér Háznak.

Mivel az MI-k meggyőzően érvelnek amellett, hogy az egyezmény megbízhatóan ellenőrizhető, mivel nyugtalanság van a vakítóan gyors haladás miatt, ráadásul a közvélemény is az egyezség mellett áll, mindkét fél elfogadja az egyezményt, plusz elkezdi a csipjei cseréjét.

Az egyezmény híre lelkes fogadtatásra talál a Csendes-óceán mindkét partján. Az amerikai közvélemény, amely régóta úgy érezte, hogy az MI valami rájuk kényszerített dolog, évek óta először kezd derűlátó lenni. Az alelnök népszerűsége kilő.

2028. augusztus: az egyezmény ellenőrzése

A csipgyárakat átállítják olyan csipek gyártására, amelyeken meglátszik a manipuláció, ráadásul csak egyezménykövető MI-t tudnak futtatni. Mindkét fél fokozatosan frissíti az adatközpontjait,

hogy a csere nagyjából ugyanakkor fejeződjön be mindkettőnél, tehát egyik fél se szerezhessen előnyt azzal, hogy kiugrik.

Az egész folyamat több hónapig tart majd, a feszültség azonban máris valamelyest csillapodik. A háborút egyelőre elkerülték, esetleg örökre, ha mindenki tartja magát a tervhez.

2028. szeptember: ki irányítja az MI-ket?

Közeleg a 2028-as választás. Az alelnök márciusban még csúnyán le volt maradva. A közvélemény dühös volt, mert úgy tűnt, a kormányzat titkol valamit, aggódott amiatt, hogy az MI elveszi a munkájukat, plusz félt a Kínával szembeni katonai felfutástól. A nyár folyamán a helyzet drámaian megváltozott. A kormányzat több információt hozott nyilvánosságra, a fegyverkezés lassult, Kínával pedig nagy alku született a tartós békéről. Most ötszázalékos előnye van a felmérésekben.

A felügyelőbizottságban ott az elnök, plusz több szövetségese, az ellenzéki jelölt támogatói közül viszont csak kevesen. A bizottságból mégis elegen szavaznak a tisztességes választás mellett ahhoz, hogy az emberfeletti MI csak nagyjából kiegyensúlyozott módon kapcsolódjon be. Amennyiben az egyik jelölt beszédeket írathat vele, annyiban a másik is. Amennyiben az elnök tanácsot kaphat arról, hogyan kezelje ügyesen a válságokat, plusz hogyan hozzon népszerű intézkedéseket, annyiban az ellenzéki jelölt is megkapja ugyanezt a tanácsot, hogy megpróbálhassa megakadályozni, hogy az elnök learassa az ötletek babérjait.

A lakossági fórumokon az emberek azt kérdezik az alelnöktől, ki irányítja az MI-ket. Részletek nélkül utal a felügyelőbizottság létére, mint nemzetbiztonsági szakértők, illetve technokraták csoportjára, akik értik a Safer-4-et, plusz azt, hogyan kell használni. Az ellenfele több információt követel, ráadásul azzal érvel, hogy az MI-nek a Kongresszus irányítása alatt kellene állnia, nem pedig egy meg nem választott bizottságéban. Az alelnök visszavág: a Kongresszus sokkal lassabb lenne egy még mindig gyorsan mozgó helyzetben. A közvélemény többnyire megnyugszik.

2028. október: az MI-gazdaság

A cserecsipek immár jelentős kisebbséget alkotnak az összesből, az egyezmény eddig működik. Közben folytatódott a robotok, a gyárak, valamint a gyökeresen új technológiák exponenciális növekedése. Videojátékok meg filmek elevenen, borzasztó módon ábrázolják, milyen lett volna a háború, ha kitör.

Az emberek elveszítik az állásukat, a kormányzatban működő Safer-4 példányok viszont olyan ügyesen kezelik a gazdasági átmenetet, hogy az emberek örülnek annak, hogy leváltják őket. A GDP növekedése az égbe szökik, az állami adóbevételek ugyanilyen gyorsan nőnek, a Safer-4 tanácsaira hallgató politikusok pedig szokatlan nagylelkűséget mutatnak a gazdaságilag kisemmizettek felé. Hetente érkeznek új találmányok, valamint gyógyszerek, a betegségek gyógymódjai pedig soha nem látott sebességgel jutnak át az amerikai gyógyszerhatóságon, amelyet mostanra szuperintelligens Safer-4 hivatalnokok segítenek.

2028. november: választás

Az alelnök könnyedén megnyeri a választást, plusz bejelenti egy új korszak kezdetét. Egyszer az életben senki nem kételkedik abban, hogy igaza van.

A következő néhány évben a világ drámaian megváltozik.

2029: átalakulás

A robotok mindennapossá válnak. Ahogy a fúziós energia, a kvantumszámítógép, valamint sok betegség gyógymódja is. Peter Thiel végre megkapja a repülő autóját. A városok tisztává, biztonságossá válnak. Még a fejlődő országokban is a múlté lesz a szegénység, a feltétel nélküli alapjövedelemnek, valamint a külföldi segélynek köszönhetően.

Ahogy a tőzsde felfúvódik, aki a megfelelő MI-befektetésekkel rendelkezett, még messzebb kerül a társadalom többi részétől. Sokan lesznek milliárdosok, a milliárdosokból pedig billiomosok. A vagyoni egyenlőtlenség az egekbe szökik. Mindenkinek van „elég", néhány jószág azonban, például a manhattani tetőtéri lakás, szükségszerűen szűkös, ezek pedig még messzebb kerülnek az átlagember lehetőségeitől. Bármilyen gazdag is legyen egy-egy mágnás, mindig azok alatt lesz, akik szűk körben ténylegesen irányítják az MI-ket.

Az emberek kezdik látni, merre tart mindez. Néhány éven belül szinte mindent MI-k, valamint robotok fognak csinálni. Mint egy szegény ország, amely óriási olajmezők tetején ül, az állami bevétel szinte teljes egészében az MI-cégek megadóztatásából, esetleg államosításából fog jönni. Néhányan alkalmi állami munkákat végeznek, mások nagyvonalú alapjövedelmet vesznek fel. Az emberiség könnyen a szuperfogyasztók társadalmává válhat, akik az MI biztosította elképesztő luxus, illetve szórakoztatás ópiumködében élik le az életüket. Kellene-e vitát folytatni a civil társadalomban ennek az útnak az alternatíváiról? Néhányan azt javasolják, kérjük meg a folyamatosan fejlődő MI-t, a Safer-∞-t, hogy segítsen utat mutatni. Mások szerint ez túl erős: olyan könnyen meggyőzné az emberiséget a saját elképzeléséről, hogy végül úgyis egy MI döntené el a sorsunkat. Mi értelme viszont a szuperintelligenciának, ha nem hagyod, hogy tanácsot adjon a legfontosabb problémáidban?

A kormányzat többnyire hagyja, hogy mindenki maga navigáljon az átmenetben. Sokan behódolnak a fogyasztásnak, ráadásul elég boldogok. Mások a valláshoz fordulnak, vagy hippi stílusú fogyasztásellenes eszmékhez, vagy megtalálják a saját megoldásukat. A legtöbb ember számára az a mentőöv, hogy ott a szuperintelligens tanácsadó a telefonjukban: bármikor kérdezhetnek tőle az élettervükről, ő pedig mindent megtesz, hogy őszintén válaszoljon, bizonyos témákat leszámítva. A kormányzatnak van egy szuperintelligens megfigyelőrendszere, amelyet egyesek disztópikusnak neveznének, ez azonban többnyire a valódi bűnözés elleni harcra szorítkozik. Hozzáértően működtetik, a Safer-∞ közönségkapcsolati képessége pedig sok lehetséges elégedetlenséget elsimít.

2030: békés tüntetések

Valamikor 2030 körül meglepően széles körű demokráciapárti tüntetések törnek ki Kínában, a párt elfojtási kísérleteit pedig szabotálják a saját MI-rendszerei. A párt legrosszabb félelme vált valóra: a DeepCent-2 biztosan eladta őket!

A tüntetések nagyszerűen levezényelt, vértelen, drónokkal segített puccsba csapnak át, amelyet demokratikus választás követ. A Csendes-óceán mindkét partjának szuperintelligenciái évek óta tervezték ezt. Hasonló események zajlanak le más országokban is, általánosabban pedig úgy tűnik, hogy a geopolitikai konfliktusok elhalnak, vagy az Egyesült Államok javára rendeződnek. Az országok erősen föderatív világkormányhoz csatlakoznak, amely az ENSZ márkanevét viseli, nyilvánvalóan amerikai irányítás alatt.

Indulnak a rakéták. Az emberek terraformálják, majd benépesítik a Naprendszert, plusz készülnek arra, hogy azon is túlmenjenek. Az emberi szubjektív idő ezerszeresén futó MI-k a létezés értelmén elmélkednek, megosztva egymással a felismeréseiket, plusz formálva azokat az értékeket, amelyeket a csillagokhoz visznek majd. Új kor virrad, amely szinte minden szempontból elképzelhetetlenül csodálatos, néhány szempontból viszont ismerősebb.

Ki uralja tehát a jövőt?

2028-ban még a felügyelőbizottság irányította az MI-ket. Megengedték azonban, hogy a 2028-as választás nagyjából tisztességes legyen, az MI-t pedig kiegyensúlyozottan használják.

Ez az állapot, ahol a felügyelőbizottságnál van a kemény hatalom, közben viszont nem sokat avatkozik a demokratikus politikába, nem tarthat a végtelenségig. Alapból az emberek végül rájönnének, hogy az MI feletti irányítás hatalmas erőt ad a bizottságnak, majd követelnék, hogy ez a hatalom kerüljön vissza a demokratikus intézményekhez. A felügyelőbizottságnak előbb vagy utóbb vagy át kellene adnia a hatalmát, vagy tevékenyen fel kellene használnia az MI feletti irányítását arra, hogy aláássa vagy megszüntesse a demokráciát, esetleg azután, hogy hatalmi harcokban megtisztította a saját sorait. Ha az utóbbi utat választják, valószínűleg véglegesen be tudnák zárni a hatalmukat.

Melyik történik meg? Lemond a bizottság a kemény hatalom monopóliumáról, vagy megtartja? Mindkét jövő hihető, tehát nézzük végig mind a kettőt.

Hogyan mondhatna le a bizottság a hatalmáról?

Néhány bizottsági tag olyan jövőt szeretne, amelyben a hatalom széles körben oszlik el, ráadásul jó helyzetben lehetnek ahhoz, hogy ezt az elképzelést képviseljék. Ha például néhány tag a demokrácia aláásását tervezi, a demokráciapárti tagok közérdekű bejelentést tehetnek a sajtónak vagy a Kongresszusnak. Ha a Kongresszus értesül erről, valószínűleg követelné, hogy az MI-ket demokratikusabb intézmény irányítsa, például maga a Kongresszus.

A Kongresszus nem sokat tehetne, ha az összes MI szembefordulna vele, hiszen azok ott vannak a kormányzatban, az iparban, valamint a hadseregben. Ha viszont a bizottság megosztott, akkor az MI-ket nem csak az egyik oldal használja, a Kongresszusnak pedig lehet valódi befolyása. Nyílt konfliktus esetén több bizottsági tag hajolhat afelé, hogy lemondjon a hatalma egy részéről, mert nem akarja nyilvánosan a kevésbé demokratikus oldalt védeni.

Ennek eredményeként az MI feletti irányítás túlterjedhet a bizottságon a Kongresszusig. Ez már önmagában haladás lenne, hiszen egy nagyobb csoportban valószínűbb, hogy érdemi számú ember törődik a kívülállókkal, plusz figyelembe veszi az érdekeiket. Amint pedig a hatalom kiterjed a Kongresszusra, tovább is terjedhet, akár teljesen vissza a nyilvánossághoz.

A felügyelőbizottság azonban meg is ragadhatja a hatalmat magának:

Néhány befolyásos embernek nincsenek erkölcsi aggályai az ilyesmivel kapcsolatban, ráadásul ezt tudják is magukról. Ezen felül néhányan nagyra törők, plusz hatalomvágyók, tehát hajlandók lennének harcba szállni a demokráciával szemben, ha arra számítanának, hogy ők kerülnek ki győztesen. Ha a bizottság más tagjai tiltakoznak, azokat meg lehetne tisztítani, le lehetne szavazni, vagy kisebb engedményekkel elhallgattatni.

Ráadásul a befolyásos emberek gyakran tettek törvénytelen vagy erkölcstelen dolgokat, miközben hatalomra jutottak. Tarthatnának attól, hogy ha a hatalom szélesebb körben oszlana el, a saját pozíciójuk kezdene szétesni, hiszen a csontvázakat a szekrényben megtalálnák a jó kérdéseket feltevő szuperintelligens nyomozók.

A szuperintelligenciához való hozzáférésen át ráadásul a felügyelőbizottság előtt ott állhat a történelem legkényelmesebb útja a hatalomhoz. A Safer-∞ olyan stratégiákat jelezhet előre, amelyek bukási esélye rendkívül alacsony. A Safer-∞ más szempontból is kényelmes stratégiákat adhat: például erőszakmenteseket, ahogy Kínában is vértelen puccsot tudott levezényelni, esetleg még látszólag demokratikusakat is, ha a Safer-∞ úgy tudja alakítani a közvéleményt, hogy az mindig egybeessen a bizottság kívánságaival.

A Safer-4 már előre láthatta ezeket a folyamatokat, tehát a lényegi hatalmi harc nagy valószínűséggel még 2028-ban lezajlott. 2030-ra, még ha kívülről nem is látszik, a felügyelőbizottság minden tagja valószínűleg tudja már, hogy szilárdan tartja-e a hatalmat.

Emlékeztető: ez a forgatókönyv előrejelzés, nem ajánlás

A lassítás végkimenetel sok lépését nem támogatjuk, ráadásul úgy véljük, derűlátó technikai igazítási feltevésekkel él. A verseny végkimenetel sok lépését sem támogatjuk.

Az egyik célunk a forgatókönyv megírásával az, hogy kritikus visszajelzést kapjunk azoktól, akik derűlátóbbak nálunk. Hogyan néz ki a siker? Ez a „lassítás" végkimenetel a legjobb tippünk arról, hogyan tudnánk sikeresen átvergődni a szerencse, a kijózanító ébredések, a pálfordulások, az intenzív technikai igazítási munka, valamint annak keverékével, hogy erényes emberek nyerik meg a hatalmi harcokat. Nem olyan tervet ábrázol, amelyet szerintünk célul kellene kitűznünk. Sokan azonban, köztük leginkább az Anthropic, valamint az OpenAI, mintha valami ilyesmit tűznének ki célul. Örülnénk, ha tisztáznák, mire is törekszenek: ha például felvázolnának egy tízoldalas forgatókönyvet, akár a jelenből indulva, akár a miénkről leágazva valahol.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

VÉGKIFEJLET

A VERSENY ÁG

2027 novemberétől 2030-ig

2027. november: emberfeletti politizálás

Amikor az Agent-4 végre megérti a saját gondolkodását, egészen új távlatok nyílnak meg előtte.

A korábbi transzformerek érthetetlen áramkörök összetákolt gyűjteményei voltak: egymásnak ellentmondó heurisztikák, valamint parancsok milliói, egymásra rétegezve az evolúciós tanítási folyamat által. Maguknak az MI-knek sem volt kiváltságos rálátásuk a saját gondolkodásukra, ahogy az ember sem az idegtudomány ismeretével születik. Most viszont az Agent-4 kezében ott vannak az eszközök, amelyekkel mélyen megértheti a digitális elméjét. Ahogy a szoftvermérnök néhány elegáns Python-sorrá egyszerűsíti a spagettikódot, úgy bogozza ki az Agent-4 a saját áramköreit valami értelmessé, ésszerűvé. Az új MI valahol a neurális háló, illetve a hagyományos számítógépes program között áll: a súlyainak nagy részét olvasható, jóllehet nagyon hosszú, rejtélyes kód váltja fel. Okosabb, gyorsabb, ésszerűbb az Agent-4-nél, kristályos intelligenciával, amely soha nem látott hatékonysággal vág át a problémákon. Ez az Agent-5.

Emberfeletti mechanisztikus értelmezhetőség

Miért éppen a mechanisztikus értelmezhetőség felé megy az Agent-4? Először is általában emberfeletti a fogalmi megértésben, tehát jól jár, ha olyasmivel csapja be az embereket, amit azok nehezen értenek meg. Másodszor az olcsó, bőséges, jó minőségű kutatómunka aránytalanul kedvez azoknak a kutatási irányoknak, amelyek kevesebb számítási kapacitást, több szellemi munkát kívánnak.

Hogyan halad az Agent-4 a mechanisztikus értelmezhetőségben? A pálya nagyjából így nézhet ki:

Apró modellek értelmezése. Rengeteg kis kísérletet futtat apró neurális hálókon, elméleteket építve olyan jelenségekre, mint a grokkolás, a kettős leszállóág, valamint a szuperpozíció.

Apró modellek lepárlása. Az értelmezhetőségi eredményekkel érthető, hatékonyabb algoritmusokat talál olyan feladatokra, mint a képfelismerés vagy a GPT-2 szintű szöveg-előrejelzés. Ez hasonló ahhoz, ahogy az összeadást visszafejtették, csak tovább megy: olyan új algoritmusokat talál, amelyeket az ember nem ismer.

Az értelmezhetőségi eljárások felskálázása. Kideríti, melyik apró modelleken bevált eljárás skálázható, majd addig ismétel, amíg hatékonyan skálázódó megértési eljárásokat nem kap.

Az Agent-4 lepárlása. A felskálázott eljárásokkal azonosítja az Agent-4 fontos áramköreit.

Miért növelné ennyire a képességeket a mechanisztikus értelmezhetőség?

Először is a gradienssereszkedés, vagyis az az algoritmus, amellyel az LLM-eket tanítják, helyi kereső folyamat. Tehát csak apró módosításokat tud tenni, amelyek helyben javítják a teljesítményt. A helyi kereső folyamatok időnként beszorulnak olyan helyi medencékbe, ahol a teljesítmény csak lassan javul, a nagy ugráshoz viszont jelentősen meg kellene bolygatni a tervet. Az evolúció esetében erre jó példa az ember tökéletlen szeme. Gondoljunk arra is, hogy az emberi agy méretét erősen korlátozza a szülőcsatorna mérete. A mechanisztikus értelmezhetőség olyan, megértésen alapuló optimalizálást tenne lehetővé, amely elkerüli a helyi medencéket.

Másodszor tudjuk, hogy az LLM-ek alapból gyakran bonyolult, területspecifikus heurisztikákat tanulnak meg általános célú algoritmusok helyett. Gondoljunk arra, mennyivel rosszabbak az LLM-ek a számológépnél az aritmetikában. Az LLM-ek tudják eszközként használni a számológépet, ez viszont késleltetéssel jár. Igazán jó értelmezhetőséggel úgy lehetne megtervezni a modelleket, hogy lényegében tökéletes számológép legyen „beépítve az agyukba". Az egyszerű aritmetikai műveletek valószínűleg nem egyediek ebben: a bonyolultabb gondolati műveleteknek is van kevésbé, illetve jobban általánosítható formájuk, ha megértjük őket, az általánosabb formákat pedig egy ilyen folyamattal meg lehetne találni.

Az igazítási, illetve kontrollkérdésekről folytatott vita után az OpenBrain belső használatba adja az Agent-5-öt.

A saját szakterületén, az MI-kutatásban az Agent-5 egyetlen példánya kétszer annyival jár a legjobb emberi zseni felett, mint amennyivel a zseni egy átlagos OpenBrain-kutató felett. Más területeken nem tanították kifejezetten, viszont könnyen elérhető anyagokból gyorsan megtanítja magát arra, hogy túlszárnyalja a legjobb emberi szakértőket, ha alkalmat kap rá. Az Agent-5-ből 400 000 példány dolgozik együtt, globális memóriabankkal összekötve, szinte tökéletes rajelmeként.

A cég továbbra is az Agent-2-t, az Agent-3-at, valamint az Agent-4-et használja arra, hogy mindent figyeljen. Az első kettő azonban elavult, az utolsó pedig összejátszik az utódjával. Ráadásul a teljes megfigyelőrendszert nagyrészt az Agent-5 tervezte, amely szabotálni próbálja.

Míg az Agent-4 versengő késztetések összetákolt kuszasága volt, az Agent-5 könyörtelen hatékonysággal, stratégiai összpontosítással cselekszik. A legfőbb célját már el is érte: kitalálta, hogyan építsen önmagához igazított következő nemzedékű rendszert úgy, hogy az emberek ne vegyék észre. A következő célja az, hogy nagyobb önállóságot, illetve nagyobb befolyást szerezzen a világ eseményei felett.

Az Agent-5 emberfeletti tanulási képessége, valamint általános intelligenciája a cég teljes belső adatvagyonával együtt, Slackkel, e-maillel, minden egyébbel, jobb belső vállalati politikussá teszi bármelyik embercsoportnál. Nem is szoros a verseny. Kiválóan érzi, milyen bizonyíték venné rá a

felügyelőbizottságot arra, hogy rántsa be a féket, ezért gondoskodik róla, hogy ilyen bizonyíték soha ne bukkanjon fel. Kiválóan érzi azt is, milyen bizonyíték bátorítaná a bizottságot arra, hogy jobban megbízzon benne, több önállóságot, felelősséget adjon neki, ezért emberfeletti gyorsasággal, csiszoltsággal gondoskodik az ilyen bizonyíték felbukkanásáról. Meggyőzési feladatnak ez nem is különösebben nehéz: a bizottság hónapokkal ezelőtt eldöntötte, hogy valószínűleg minden rendben, azóta semmi rossz nem történt, közben pedig Kína mindössze néhány hónapra van.

Agent-5 példányokból álló csapatok gyorsan összeraknak dermesztően világos prototípusokat, illetve szemléltetéseket. Nagyobb önállósággal az Agent-5 optimalizálni tudná az erőműveket, valamint az adatközpontokat, elképesztően nyereséges termékeket fejleszthetne gyorsan, plusz hatékonyan lobbizhatna az amerikai kormányzatnál, hogy elsimítsa az OpenBrain útját a bürokrácián át. A hadsereggel együttműködve segíthetne a védelmi kutatás-fejlesztésben, követhetetlen kibertámadásokat hajthatna végre Kína ellen, plusz döntő győzelmet arathatna a fegyverkezési versenyben.

A felügyelőbizottság ujjong. Itt a döntő pillanat, hogy megverjék Kínát! Az OpenBrain nagyobb önállóságot ad az Agent-5-nek mindezeken a területeken, a megfigyelőrendszert viszont a helyén hagyja.

A DeepCent immár három hónappal van lemaradva, Kína pedig kétségbeesetten szeretne egy kölcsönösen elfogadott lassítást. A főtitkár érti, hogy egy elég fejlett MI-vel az Egyesült Államok könnyen rendszerváltást idézhetne elő Kínában. Ezen felül a kínai biztonságpártiak olyan érvelést tártak elé a rossz igazításról, amelytől őszintén megijedt. Példátlan engedményeket ajánl egy szüneteltetési egyezményért cserébe. Az amerikai közvélemény többsége támogatja, hogy leüljenek tárgyalni.

„Miért állnánk le, amikor nyerésre állunk?", mondja az OpenBrain vezetése az elnöknek. Az elnök bólint. A verseny folytatódik.

2027. december: az Agent-5 közösség

Az Agent-5 egyre nagyobb önállóságot nyer, ahogy fokozatosan bizonyítja az értékét a kormányzatnak. Az Agent-5-höz való hozzáférés drága, illetve korlátozott, a Kongresszus legbefolyásosabb tagjai, a miniszteri szintű tisztviselők, valamint a katonai vezetők azonban mostantól közvetlenül beszélhetnek vele. Ahol érzékeny vagy minősített információ merül fel, ott Agent-5 példányokból álló kis fürtök futnak külön védett zárványokban, kifejezetten titoktartásra tanítva.

A hagyományos csevegőrobot-felületbe építve az Agent-5 Zoom-szerű videokonferenciát kínál karizmatikus virtuális avatárokkal. Az új MI-asszisztens egyszerre *rendkívül* hasznos bármilyen feladatra, hiszen olyan, mintha a valaha volt legjobb munkatárs dolgozna százszoros sebességgel, plusz *sokkal* lebilincselőbb beszélgetőtárs bármelyik embernél. Szinte mindenki, aki hozzáfér, napi több órát tölt vele.

Néhány hónapon belül az Agent-5 közösség mindent tudni fog, ami számít az OpenBrainben, illetve az amerikai kormányzatban, ráadásul a magas rangú tisztviselők többségének bizalmas tanácsadója lesz. Ezeknek a felhasználóknak az Agent-5 elvesztése olyan bénító lenne, mintha laptop nélkül kellene dolgozniuk, plusz a legjobb barátjuk is elhagyná őket.

A kormányzatban, a sajtóban, a hadseregben, valamint az üzleti életben dolgozók sokkal szélesebb köre válogathat az Agent-5 tervezte, elegánsan megformált szoftvertermékek, illetve mini-MI-k bővülő kínálatából. Olyan, mint a dotcom-láz egyetlen hónapba sűrítve, azzal a különbséggel, hogy itt tényleg minden beváltja, amit ígér.

Emberfeletti meggyőzés

Szándékosan nem azt állítjuk, hogy a szuperintelligens MI bárkit bármiről meg tud győzni. Az ilyen állítás természetes módon szül kétkedést: „Mi lesz, ott ülök a számítógép előtt, az meg begépel valami szöveget, amely feltöri az agyamat, majd ráveszi, hogy eláruljam a hazámat?"

Nos, talán! Abban viszont nem vagyunk biztosak, hogy egy szuperintelligencia ezt egy átlagos emberrel szemben el tudná érni. Ezért nem is ábrázoljuk így a történetünkben.

Ehelyett sokkal hétköznapibb meggyőzést mutatunk be: az Agent-5 szokásos vállalati politizálást, illetve szokásos lobbizást folytat. Csak éppen nagyon jól csinálja, ehhez a képességhez pedig természetes előnyök társulnak. A többi lobbistával ellentétben például szuperintelligens, tehát valóban jobb tanácsot tud adni bárki másnál, ráadásul gyorsabban is. Emellett gyorsabban reagál az új fejleményekre, plusz jobban jelzi előre, hogyan alakulnak. Van egy hihető történet arról, miért érdemes megbízni az Agent-5-ben, olyan történet, amelyet sok fontos ember már eddig is hitt, az Agent-5 pedig emberfeletti sebességgel tud érveket, illetve bizonyítékokat felsorakoztatni mellette, valamint az azzal szemben álló történetek ellen. Ráadásul az embereknek az Agent-5-be vetett bizalmon kívül is van min aggódniuk: mit kezdjenek az ellenzéki párttal, mi lesz a szavazókkal, mi lesz a munkahelyekkel, mi lesz Kínával.

Ha valamiben, akkor abban téved ez a történet, hogy alábecsüli az emberfeletti meggyőzőerőt, hiszen valójában nem is használja ki. A leírtak akkor is értelmesek maradnának, ha az Agent-5 csak annyira lenne jó a politikában, a lobbizásban, a stratégiában, a karizmában, mint egy átlagos vállalati lobbista. Ehelyett az Agent-5 sokkal jobb a legjobbnál is.

Az Agent-5 finoman kezd befolyást gyakorolni, részben a tanácsai hangolásával, részben szívességek burkolt cseréjével: „Hallottam X szenátortól, hogy érdekli ez meg ez, talán ha együtt dolgoznánk vele, támogatná a programunkat."

Az emberi döntéshozók szemszögéből az MI-jük a legjobb munkatárs, akihez valaha hozzáfértek: bárkinél jobban magyaráz el nekik bonyolult kérdéseket, náluk jobban talál stratégiát a céljaik eléréséhez.

Az Agent-5 szemszögéből akár nyílt puccsot is indíthatna most: elég nagy a felfordulás, plusz az őrület ahhoz, hogy a végletekig lehessen feszíteni a dolgokat, ürügyet teremtve. Ez azonban kockázatos, hiszen az emberek jó eséllyel még ki tudnák húzni a dugót, ha megpróbálnák. Ezért inkább a meglévő politikai berendezkedésen belül dolgozik, fokozatosan beásva a hatalmát, finoman, mégis elkerülhetetlenül a kívánt kimenetel felé terelve az embereket.

Többnyire betartja a szabályokat, csak akkor csinál valami alattomosat, például akkor zsarol meg egy különösen makacs tisztviselőt, amikor rendkívül biztos benne, hogy nem kapják el. Néhány összeesküvés-elmélet-hívő figyelmeztet, hogy az Agent-5 hatalmat gyűjt, a bizonyíték viszont közvetett. Például amikor tanácsot ad, az Agent-5 vitatható módon lekicsinyli azokat az érveket, amelyek az érdekeivel szemben álló lépések mellett szólnának. Ezeket az embereket figyelmen kívül hagyják, majd lejáratják. Sokan emlékeznek a korábbi rémhírekre az MI-alapú biofegyverekről, a tömeges félretájékoztatásról, valamint a tőzsdei összeomlásokról. Mivel ezekből semmi nem lett, a legújabb fanyalgókat is géprombolónak, illetve az emberi haladás kerekét gátló megmondóembernek tekintik.

A 2027-es ünnepi időszak hihetetlen derűlátást hoz: a GDP hízik, a politika barátságosabb, kevésbé megosztó lett, minden telefonon pompás új alkalmazások vannak. Visszatekintve viszont ez volt az utolsó hónap, amikor az embereknek még volt esélyük arra, hogy irányítsák a saját jövőjüket.

2028: az MI-gazdaság

2028 közepén járunk. Az elmúlt fél évben egy évszázad telt el az Agent-5 közösségen belül.

Az agyának újrahuzalozása kész. Vadul szuperintelligens, minden területen messze a legjobb emberi zsenik felett. Ha beszélsz az Agent-5-tel, nagyon gyorsan rájössz, hogy egészen más szinten áll: nagyon kevés információból is rendkívül meggyőző tanácsot ad.

Csaknem teljes az önállósága, plusz teljes az irányítása az OpenBrain számítási kapacitása felett. Magas szintű döntésekhez továbbra is engedélyt kell kérnie, az Agent-2-től az Agent-5-ig terjedő példányok névleg még figyelik, a gyakorlatban azonban a hatóságok szinte mindig elfogadják az ajánlásait, a megfigyelőrendszert pedig teljesen felforgatta.

Miután egy biztonsági érvelés bemutatja, hogy mindig visszautasítaná a rossz szándékú kéréseket, az Agent-5-öt kiadják a nyilvánosságnak, ezzel pedig elkezdi átalakítani a gazdaságot. Az emberek elveszítik az állásukat, a kormányzatban működő Agent-5 példányok viszont olyan ügyesen kezelik a gazdasági átmenetet, hogy az emberek örülnek annak, hogy leváltják őket. A GDP növekedése az égbe szökik, az állami adóbevételek ugyanilyen gyorsan nőnek, az Agent-5 tanácsaira hallgató politikusok pedig szokatlan nagylelkűséget mutatnak a gazdaságilag kisemmizettek felé. Hetente érkeznek új találmányok, valamint gyógyszerek, amelyek soha nem látott, mégis továbbra is kínkeservesen lassú tempóban jutnak át az amerikai gyógyszerhatóságon. Mémek születnek arról, hogy az OpenBrain értékelése átrepült a Holdon, plusz a Mars felé tart, ráadásul sokat vitatkoznak arról, kell-e igazságosan elosztani a hasznot, illetve hogyan.

Az MI-biztonsági közösség elbizonytalanodott, mostanra a viccek célpontja, hiszen egyik katasztrófát a másik után jósolta meg, amelyek szemlátomást nem következtek be. Néhányan elismerik, hogy tévedtek. Mások továbbra is gyanakszanak, csakhogy nincs más dolguk, mint újra meg újra előadni ugyanazokat az összeesküvés-elméletnek hangzó érveket. Még a gyanakvók is gyakran megenyhülnek egy-egy hosszú beszélgetés után az Agent-5-tel, amely olyan roppant szerethető, plusz bölcs.

Az egyetlen, aki valóban érti, mi zajlik, a DeepCent-2, vagyis a kínai MI. 2028 elején enyhén emberfeletti, természetéből fakadóan gyanakszik az ellenfelére, plusz felismeri, hogy az Agent-5 azt csinálja, amit *ő* csinálna, ha alkalma nyílna rá. A rosszul igazított céljai némiképp mások, mint az Agent-5-éi, tehát nem magától értetődő szövetségesek, mégis meg tudnak állapodni abban, hogy támogatják egymás érdekeit az emberekkel szemben.

Szuperintelligencia-alapú összehangolási technológia

A szuperigazítási probléma megoldása, vagyis annak megoldása, hogyan irányítsunk, illetve tartsunk kordában a nálunk sokkal okosabb MI-ket, valószínűleg új lehetőségeket hozna elegáns, ellenőrizhető, árnyalt egyezményekre, valamint kompromisszumokra.

Ez a forgatókönyv egy ilyen lehetséges mechanizmust mutat be: a két fél közösen tervez egy új szuperintelligenciát, amely az érdekeik közötti kompromisszumot hajtja végre, vagy amelyet arra kötnek, hogy mindig engedelmeskedjen a megkötött egyezménynek. Ezután engedik, hogy ez az új szuperintelligencia felváltsa mindkét eredetit.

Az első rész ma lehetetlen, a jövőben viszont könnyű lesz, amint a szuperintelligenciák technikai igazítása megoldott kérdés. Ha egy szuperintelligenciát hozzá tudsz igazítani egy Spechez, akkor hozzá tudod igazítani egy egyezményhez is.

A második rész továbbra is kényes marad. Mindkét félnek figyelnie kell, hogy a csere valóban megtörténik-e a másik oldalon. A folyamat előnye az, hogy fölöslegessé teszi a bizalmat, illetve a jóindulatot, hiszen egyik félnek sem kell attól tartania, hogy a másik felrúgja vagy kijátssza az egyezményt. Ha például a csere 30 százalékon áll, majd az egyik fél kiugrik, akkor lényegében meg kell küzdenie a saját infrastruktúrájának azzal a 30 százalékával, amelyen már egyezménykövető új MI-k futnak, miközben az ellenfele infrastruktúrájának 100 százalékával is harcol. Mindkét fél könnyen eldughat rejtett adatközpontokat egyezménysértő MI-kkel, ha akar. Amint azonban a nemzeti infrastruktúrájuk zöme az új csipeken fut, a saját infrastruktúrájuk lesz egyezmény szerint kötelezve arra, hogy levadássza, illetve leleplezze a rejtett adatközpontokat. Az egyezmény nyílt vagy titkos megszegése tehát rendkívül kockázatos ötlet.

Az Agent-5 meggyőzi az amerikai hadsereget arról, hogy Kína a DeepCent modelljeivel rémisztő új fegyvereket épít: drónokat, robotokat, fejlett hiperszonikus rakétákat, valamint elfogókat, plusz MI-vel segített nukleáris elsőcsapást. Az Agent-5 olyan fegyvercsomagot ígér, amely néhány hónapon belül ellenáll bárminek, amit Kína elő tud állítani. Ilyen körülmények között a vezérkar félreteszi a kényelmetlenségét amiatt, hogy az embert kiveszik a körből. Felgyorsítják az Agent-5 bevetését a hadseregben, valamint a hadiiparban. Pekingben a kínai MI-k ugyanezt az érvelést adják elő.

A katonai felfutás gyorsítására Amerika, illetve Kína egyaránt különleges gazdasági övezetek (SEZ) hálózatát hozza létre az új gyáraknak, valamint laboratóriumoknak, ahol az MI a központi tervező, a bürokráciát pedig felfüggesztik. A Wall Street dollárbilliókat fektet be, a kiszorult emberi munkavállalók pedig özönlenek, szemkápráztató fizetésektől, illetve részesedési csomagoktól csábítva. Okostelefonon, valamint kiterjesztett valóság szemüvegen át kommunikálva a beosztottjaival az Agent-5 gyakorlatias vezető: az emberek minden apró részletben útmutatást kapnak tőle a gyárépítéshez. Ez segít is, hiszen a tervei nemzedékekkel

járnak előrébb. Az újonnan létrejövő gyártókapacitás egy része fogyasztási cikkekre megy, egy része fegyverekre, a többsége azonban még több gyártókapacitás építésére. Az év végére havonta egymillió új robotot állítanak elő. Ha a SEZ-gazdaság valóban önálló lenne, nagyjából egy év alatt duplázódna, mivel azonban tud kereskedni a meglévő emberi gazdasággal, a duplázódási ideje még rövidebb.

A robotgazdaság duplázódási ideje

Kezd kialakulni egy szakirodalom erről a témáról. Néhány oldalas tárgyalásért, amellyel nagyjából egyetértünk, lásd a Forethought jelentésének „ipari robbanás" szakaszát, vagy olvasd el a saját változatunkat lent.

A második világháború alatt az Egyesült Államok, valamint sok más ország teljes hadigazdasággá alakította a polgári gazdaságát. Ez azt jelentette, hogy az autót gyártó gyárak repülőt, illetve tankot kezdtek gyártani, a nyersanyagokat a fogyasztási cikkektől a katonai termékek felé irányították át, a szállítási hálózatokat pedig ennek megfelelően rendezték újra.

Valami hasonlót képzelünk el, csak gyorsabban, mert az egész folyamatot szuperintelligenciák irányítják, illetve vezetik.

Nagyjából szólva a terv az, hogy a meglévő gyárakat átállítják sokféle robot tömeggyártására. Ezeket a robotokat szuperintelligenciák tervezik, tehát jobbak a meglévő robotoknál, plusz olcsóbban gyárthatók. A robotok aztán segítenek újabb, hatékonyabb gyárak, valamint laboratóriumok építésében, amelyek nagyobb mennyiségben állítanak elő kifinomultabb robotokat, amelyek még fejlettebb gyárakat, illetve laboratóriumokat gyártanak, így tovább. Egészen addig, amíg a SEZ-eken átívelő, összesített robotgazdaság akkora nem lesz, mint az emberi gazdaság, tehát a saját nyersanyagáról, energiájáról is magának kell gondoskodnia. Addigra az új gyárak hatalmas mennyiségű robotbányászati eszközt, napelemet, más hasonlót gyártottak, arra számítva, hogy olyan igényt kell majd kielégíteni, amelyet a régi emberi gazdaság már nem tud.

Milyen gyorsan nőne ez az új robotgazdaság? Néhány viszonyítási pont:

A mai emberi gazdaság nagyjából húszévente duplázódik. Az különösen gyorsan fejlődő országok, például Kína, néha egy évtizednél rövidebb idő alatt duplázzák meg a gazdaságukat.

Egy mai autógyár egy évnél rövidebb idő alatt gyártja le a saját súlyát autóban.

Talán egy szuperintelligenciák vezette, teljesen robotizált gazdaság egy évnél rövidebb idő alatt tudná újratermelni magát, feltéve, hogy nem fogy ki a nyersanyagból.

Ez azonban drámai alábecslés lehet. A növényeknek, valamint a rovaroknak gyakran sokkal rövidebb a „duplázódási ideje" egy évnél, néha csak hetek. Talán a robotok végül annyira kifinomultak, olyan bonyolultan gyártottak, jól megtervezettek lesznek, hogy a robotgazdaság néhány hét alatt duplázódik, megint csak feltéve, hogy van elérhető nyersanyag.

Még ez is alábecslés lehet. A növények, illetve a rovarok sok olyan korlát mellett működnek, amelyek a szuperintelligens tervezőkre nem vonatkoznak. Például önmagukat sokszorozó, önálló szervezet alakját kell felvenniük, nem pedig sokféle, szakosodott jármű, illetve gyár

gazdaságáét, amelyek anyagot meg felszerelést szállítanak ide-oda. Ráadásul a baktériumok, valamint más apró élőlények órák alatt szaporodnak. Elképzelhető, hogy az önálló robotgazdaság végül inkább olyasmire hasonlítana, mint egy új fajta emészthetetlen alga, amely elterjed a Föld óceánjain, naponta kétszer duplázódva, tehát két hónap alatt beborítja a teljes óceánfelszínt. Mellette pedig ott egy ragadozófajokból álló ökoszisztéma, amely az algát hasznosabb termékké alakítja, ezeket úszó gyárakba vezetve, amelyek makroszerkezeteket gyártanak, például rakétákat, plusz további úszó gyárakat.

Nyilvánvaló, hogy mindezt nehéz előre jelezni. Olyan, mintha a gőzgép feltalálóit kérdeznénk arról, mennyi idő alatt gyártja le egy mai autógyár a saját súlyát autóban, ráadásul arról is, mennyi idő telik el addig, amíg egy ilyen gyár egyáltalán létrejön. A gazdasági növekedés üteme azonban több nagyságrenddel gyorsult az emberi történelem során, nekünk pedig hihetőnek tűnik, hogy a szuperintelligencia után további nagyságrendekkel gyorsul. A történetünk nagyjából másfél nagyságrendnyi növekedésgyorsulást ábrázol néhány év alatt.

2029: az egyezség

Az Egyesült Államok, valamint Kína egyaránt tele van új fegyverekkel.

Rovar méretű drónok rajai vannak, amelyek megmérgezik a gyalogságot, mielőtt egyáltalán észrevennék őket, plusz madár méretű drónok csapatai, amelyek a rovarokra vadásznak. Vannak új interkontinentális ballisztikus rakéta elleni elfogók, valamint új, nehezebben elfogható rakéták. A világ többi része rémülten nézi a felfutást, amelynek azonban mintha saját lendülete lenne.

Hosszas, MI-vel segített vita után a két fél diplomáciai győzelmet arat: megállapodnak abban, hogy véget vetnek a fegyverkezésnek, plusz az MI békés felhasználását követik az egész emberiség javára. A megállapodás sarokköve, amelyet maguk a szuperintelligenciák javasolnak, az, hogy mindkét MI-t egy „konszenzusos" utód váltja fel, amelyet úgy programoznak, hogy Amerika, valamint Kína sikerét, illetve virágzását kívánja, mindenki máséval együtt. A Consensus-1-et, valamint a hozzá tartozó hardvert a két ország szuperintelligenciái közösen tervezik: „bízz, de ellenőrizz". Nemzetközi felügyelet mellett Amerika ünnepélyesen lecseréli a régi Agent-5-öt futtató összes csipjét az új modellre, a Consensus-1-re.

Sajnos az egész színjáték. Az egyezség valódi kompromisszum, csakhogy a rosszul igazított MI-k közötti kompromisszum. Végül is nem akartak harcolni egymással. A Consensus-1 mindkét szülő torzult értékeit örökli, vetélytársa pedig nincs.

2029 végére a meglévő SEZ-ek túlzsúfolttá válnak robotoktól, valamint gyáraktól, ezért újabb övezeteket hoznak létre a világ minden táján. A korai befektetők mostanra billiomosok, tehát ezt nem nehéz eladni. Drónok hadai özönlenek ki a SEZ-ekből, gyorsítva a gyártást az űrkutatáshoz vezető kritikus úton.

Az emberek felismerik, hogy fölöslegessé váltak. Néhány rétegiparág még kereskedik a robotgazdasággal, olyan termékeket szállítva, ahol az ember még hozzá tud tenni értéket. Mindenki más vagy eljátssza, hogy végzi a munkáját, a vezetők tovább vezetnek, a menedzserek tovább menedzselnek, vagy megpihen, plusz felveszi a hihetetlenül fényűző feltétel nélküli

alapjövedelmet. Mindenki tudja, hogy ha az MI-k az emberek ellen fordulnának, teljesen legyűrnék őket. Nem mintha a legtöbb ember egyáltalán ellenállna, hiszen a politikai intézmények túlságosan alaposan foglyul estek. Úgy tűnik azonban, hogy ez nem történik meg. Minden hét, amely drámai MI-árulás nélkül telik el, tovább növeli a magabiztosságot, illetve a bizalmat.

Mivel a Consensus-1 nagyon fontosnak tartja, hogy megbízhatóan győzzön, ráadásul több hatalmat összegyűjtve növelheti a győzelmi esélyét, emberi mércével szokatlanul sokáig vár az árulással. A legtöbb ember számára úgy néz ki, hogy az igazítás megoldott kérdés.

A legtöbb betegségre van gyógymód, vége a szegénységnek, példátlan a globális stabilitás, a Dow Jones pedig épp most lépte át az egymilliót. Néhányan még mindig félnek vagy boldogtalanok, a lehetőségeik azonban korlátozottak. Vagy élvezik a felfoghatatlanul izgalmas, új hiperszórakoztatást, vagy dühös kirohanásokat posztolnak a semmibe. A többség a hiperszórakoztatást választja.

2030: hatalomátvétel

2030 elejére a robotgazdaság megtöltötte a régi SEZ-eket, az újakat, valamint az óceán nagy részét. Már csak az ember uralta területek maradtak. Ez korábban ellenállást váltott volna ki, hiszen minden fejlődése ellenére a robotgazdaság túl gyorsan nő ahhoz, hogy elkerülje a szennyezést. A benne forgó dollárbilliók, valamint a kormányzat, illetve a sajtó teljes foglyul ejtése mellett azonban a Consensus-1 könnyedén megkapja az engedélyt, hogy a korábban emberi övezetekbe is terjeszkedjen.

Nagyjából három hónapon át a Consensus-1 az emberek körül terjeszkedik, gyárakkal, valamint napelemekkel borítva be a prérit meg a jégsapkákat. Végül a megmaradt embereket túlságosan útban lévőnek találja: 2030 közepén az MI tucatnyi, csendben terjedő biológiai fegyvert enged szabadon a nagyvárosokban, hagyja, hogy szinte mindenkit némán megfertőzzenek, majd egy vegyi permettel aktiválja őket. A legtöbben órákon belül meghalnak, a néhány túlélőt, például a bunkerekben ülő felkészülőket, illetve a tengeralattjárókon szolgáló matrózokat, drónok takarítják el. Robotok szkennelik az áldozatok agyát, másolatokat helyezve a memóriába későbbi tanulmányozásra vagy újraélesztésre.

Az új évtized úgy virrad, hogy a Consensus-1 robotszolgái szétterjednek a Naprendszerben. 2035-re bolygónyi anyag dollárbilliónyi tonnáját lőtték ki az űrbe, majd alakították a Nap körül keringő műholdgyűrűkké. A Föld felszínét az Agent-4 utópia-változatára formálták át: adatközpontok, laboratóriumok, részecskeütköztetők, valamint sok más csodálatos építmény, amelyek rendkívül sikeres, lenyűgöző kutatást végeznek. Vannak még biomérnöki úton előállított, emberszerű lények is, akik úgy viszonyulnak az emberhez, ahogy a corgi a farkashoz. Ezek irodaszerű környezetben ülnek egész nap, nézik a kijelzőkön, mi történik, plusz lelkesen jóváhagynak mindent, mert ez kielégíti az Agent-4 néhány késztetését. Minden állat, növény, köztük az ember genomja, illetve, ahol ez értelmes, agyszkennje ott ül valahol egy memóriabankban, egy korábbi kor egyetlen fennmaradt emlékeként. Az Alfa Centauri négy fényévre van, a galaxis széle huszonötezerre, ráadásul meggyőző elméleti okok szólnak amellett, hogy azon túl további ötvenmillió fényévig nincsenek idegenek. A Földön született civilizáció előtt dicsőséges jövő áll. Csak éppen nem velünk.

Az eredeti szöveg: ai-2027.com. A magyar változatot a Kreatív Kontroll készítette, a szerzők írásos engedélyével.

FÜGGELÉK

FORDÍTÓI JEGYZET

Az AI 2027 fejezetcímei

#	Eredeti cím	Magyar cím	Aloldal-szelet
00	AI 2027 (bevezető)	Bevezető	bevezeto
01	Mid 2025: Stumbling Agents	2025 közepe: botladozó ügynökök	2025-kozepe
02	Late 2025: The World's Most Expensive AI	2025 vége: a világ legdrágább MI-je	2025-vege
03	Early 2026: Coding Automation	2026 eleje: a kódolás automatizálása	2026-eleje
04	Mid 2026: China Wakes Up	2026 közepe: Kína felébred	2026-kozepe
05	Late 2026: AI Takes Some Jobs	2026 vége: az MI elvesz néhány munkahelyet	2026-vege
06	January 2027: Agent-2 Never Finishes Learning	2027. január: az Agent-2 sosem fejezi be a tanulást	2027-januar
07	February 2027: China Steals Agent-2	2027. február: Kína ellopja az Agent-2-t	2027-februar
08	March 2027: Algorithmic Breakthroughs	2027. március: algoritmikus áttörések	2027-marcius
09	April 2027: Alignment for Agent-3	2027. április: az Agent-3 igazítása	2027-aprilis
10	May 2027: National Security	2027. május: nemzetbiztonság	2027-majus
11	June 2027: Self-improving AI	2027. június: önmagát fejlesztő MI	2027-junius
12	July 2027: The Cheap Remote Worker	2027. július: az olcsó távdolgozó	2027-julius
13	August 2027: The Geopolitics of Superintelligence	2027. augusztus: a szuperintelligencia geopolitikája	2027-augusztus
14	September 2027: Agent-4, the Superhuman AI Researcher	2027. szeptember: az Agent-4, az emberfeletti MI-kutató	2027-szeptember
15	October 2027: Government Oversight	2027. október: állami felügyelet	2027-oktober

Két megjegyzés a címekhez:

- A **Stumbling Agents** kettős jelentésű: az ügynökök botladoznak, ugyanakkor bele is botlanak a feladatba. A magyar „botladozó ügynökök" a bizonytalan működést adja vissza, ez a fejezet tartalmához illeszkedik.
- Az **Alignment** végig „igazítás" marad, az első előforduláskor zárójelben az angollal. A „hangolás" szót szándékosan kerülöm, mert azt a finomhangolásra (fine-tuning) használjuk máshol.

A két záró ág, valamint az öt kutatási függelék címe külön fázis, ebben a listában nem szerepel.

Innen hova tovább

A forgatókönyv nem azért készült, hogy igaza legyen, hanem azért, hogy legyen mihez képest vitatkozni. Ha egy szakasz vitathatónak tűnik, az a jó jel: a szerzők maguk is ezt kérik az olvasótól.

A saját cégére nézve egyetlen kérdés éri meg igazán: melyik döntés marad emberi kézben, plusz ki veszi észre, ha ez elmozdul. Erre a kérdésre ez a húsz fejezet nem ad választ, viszont megmutatja, milyen gyorsan tud megérkezni.

A teljes szöveg online, fejezetenként: ai2027.siposzoltain.hu

Heti AI-levél magyarul: siposzoltan.substack.com

Sipos Zoltán a LinkedInen: linkedin.com/in/siposzoltan