Nyitó gondolatok a Mesterséges Intelligencia kapcsán

Ez az anyag már sokadik változatát éli, többször is újrakezdtem egy-egy témát, mivel túlságosan hosszúra nyúltak az egyes szakaszok. Ezért néhány saját gondolat kifejtése mellett inkább igyekszem azokat a fogalmakat összegyűjteni, amelyekkel találkoztam a témában, és amelyek ismerete nélkül szerintem nagyon nehéz helyes hozzáállást kialakítani a témához.

Lévén én is éppen csak ismerkedem a témával, a lista nem feltétlenül teljes, és lehet, hogy néhány súlyosabb kérdés kimarad ebből az anyagból. Remélem, idővel teljesebb és koherensebb képet fogok tudni kialakítani a témában...

Néhány általános gondolat

Az M.I. (Mesterséges Intelligencia, vagy angolul Artificial Intelligence, A.I.) kapcsán van néhány általános gondolat bennem, ami számomra a keretét adja annak, ahogyan jelenleg gondolkodom a dologról. Ismét megjegyzem, az elgondolásaim még rendezetlenek, illetve leginkább hiányosak (lehetnek). Még sokat kellene tanulnom róla, hogy egy egységes képet tudjak festeni...

Mindjárt az elején szeretnék egy dolgot kiemelni. Én nem azt tartom félelmetesnek, hogy egyszer az A.I. majd "öntudatra ébred", mint a 80-90-es évek sci-fi történeteiben. Ugyanis ennek a mondatnak nem sok értelme van (amíg nem tudjuk tudományos igényességgel megmondani, hogy mit jelent az "öntudat", vagy az “öntudatra ébredés”).

Az A.I. létrehozásában és fejlesztésében a lényegi cél, hogy az emberi gondolkodás egyes részeit bízzuk rá, mivel gyorsabb, nagyobb erőforrásokat tud mozgósítani (például számítási kapacitás, adat-szintű memória, stb.). És számomra éppen ebben van a dolog (technológiai oldalról) veszélye: **az emberi gondolkodás hibáival (is) felruházunk egy olyan rendszert, amely bizonyos irányokba számunkra elképzelhetetlen erőforrásokat képesek mozgósítani**.

A használat oldaláról pedig az a gondom, hogy az elmúlt időszak nagy technológiai-szociológiai kihívásainak jó részét bebuktuk. Hosszú felsorolás lenne mindent összeszedni, de csak egy példa: a mobil-technológia és a hozzá tartozó applikációk kezelése társadalmi szinten nem jutott el a felelősségteljes, felnőtt állapotba. A legtöbben igazából már nem is foglalkozunk a helyes használat kérdéseivel (veszélyek elhárítása/kezelése, illetve a sikeres használat módozatai), beépült a mindennapjainkba, és az ezzel a technológiával együtt járó következmények mélységeibe folyamatosan csúszunk bele, immáron ellenállás nélkül.

Az ember viszonyulása az M.I.-hez

Az M.I.-vel kapcsolatban háromféle viszonyulást tudtam felfedezni, ami nem pusztán elutasítja annak használatát (plusz egy negyediket, ami félelmetes). Mindhárom (négy) más és más szabályokat jelent számunkra. Külön nehézség, hogy ezekkel a szabályokkal nagyon keveset foglalkozunk, vagy legalábbis a köztudatban nem igazán ismert elég részletesen.

Ez engem személy szerint azért is érdekel, mert általában felnőtt egyénként a legtöbb egyszerű technikai eszköz esetén elég jól ismerjük a legalapvetőbb veszélyeket, azok elkerülésének módjait, a következményekkel való megküzdést - ahogy ismerjük a jó működtetés alapelveit és hogyan-ját is. Hasonló igaz az emberi kapcsolatainkra is - leglábbis elméletben.

Az A.I. tekintetében annak ellenzői gyakran túlzó, vagy félrevezető veszélyekre hívják fel a figyelmet (gyakran egyébként nem alaptalanul), míg az ezen technológiát felkarolók jó része a veszélyekkel alig foglalkozik, vagy nagyon felszínesen, közhelyesen (már azokban a forrásokban, amikkel találkozom). Igazán érett elgondolásokkal leggyakrabban olyan kutatók oldalról láttam, akik nem az A.I. rendszerek finomhangolásával, illetve alkalmazások fejlesztésével foglalkoznak, hanem az A.I. működésének alapjait kutatják (persze ott is vannak “rózsaszín felhőbe burkolt” gondolkodók).

Node lássuk az említett három megközelítést!
1. Az A.I. egy eszköz.
2. Az A.I. az ember társa/partnere (bizonyos feladatok elvégzésénél, idővel majd általánosságban is)
3. Az A.I. az ember helyettesítője (bizonyos feladatok elvégzésénél, idővel majd általánosságban is)

Mindegyik megközelítés nagyon máshová vezet, és erősen függ attól, hogy ki, mire, és milyen A.I.-t használ.

Az már veszélyesebb dolog, ha valaki az A.I.-t az ember egyfajta továbbfejlesztéseként, vagy a "törzsfejlődés" következő szakaszaként kezeli - ez a negyedik hozzáállás.

A fenti három megközelítés külön-külön elgondolásokat igényel, és mindenki, aki A.I.-t használ, eldöntheti, hogy miként teszi azt - ezzel választva bizonyos előnyöket és veszélyforrásokat, illetve szabályokat egyaránt. De emellett fontos az is, hogy lássuk, mások hogyan viszonyulnak hozzá, mert ez erősen befolyásolhatja a mi életünket is.

Megjegyzem, hogy a fenti megközelítések más-más súllyal, de a többi technikai eszközünk esetén is jelen vannak/lehetnek, de talán az A.I. az első olyan technológia, amit már kifejezetten ezekkel a hozzáállásokkal, ezekre a célokra fejlesztenek, kimutatva a korábbi "eszköz" formulán.

Van két fogalom-pár, amire érdemes figyelemni.

- Specializált Mesterséges Intelligencia (Artificial Narrow Intelligence - ANI): adott konkrét feladat(ok) elvégzésére, adott kérdés(ek) megválaszolására fejlesztett M.I.
- Általános Mesterséges Intelligencia (Artificial General Intelligence - AGI): az emberi gondolkodásmód egészét másolni képes mesterséges intelligencia

Sokan ez utóbbit tartják veszélyesnek, azt vízionálva, hogy majd átveszi az emberiség felett az irányítást. A kevésbé erős megközelítésben egyszerűen csak a meghatározó tevékenységekben az AGI átveszi majd az ember helyét, hiszen majdnem minden emberi feladatot gyorsabban, hatékonyabban, megbízhatóbban (no, erre még majd visszatérünk) végez majd el.

Bár a több feladatra is alkalmas A.I.-okat szeretik Általános Mesterséges Intelligenciának hívni, azok nem azok. Az AGI létrehozására a vezető kutatók szerint még évtizedekre van szükség.

A másik fogalompár (ami mellé egy harmadikat is beteszek) közelebb áll a jelenkorhoz, és magában hordozza a közvetlen veszélyt.
- Generatív Mesterséges Intelligencia (Generative A.I.): adott tartalmakat létrehozó (generáló) M.I. - szövegek, képek, hang, kód, stb.
- Ügynökalapú (szerintem naaaagyon rossz a magyar fordítás ehhez, egyébként Agentic A.I.): célokat fogalmaz meg, tervez, és minimális emberi behatás mellett döntéseket hoz és cselekszik.

Az eszköz vagy partner kérdés főként ez utóbbinál izgalmas kérdés, de a használatának veszélyeinek többsége eléggé nyilvánvaló.

De talán a legfontosabbak nem is annyira nyilvánvalóak. Kutatók észrevették, hogy a fejlesztés alatt álló agentic A.I.-ok önfenntartást célzó döntéseket (is) hoznak, ráadásul “előre megfontolt szándékkal” hazudnak is - a felhasználók az A.I. döntéshozatali folyamatát nem látják, de a kutatóknak meg vannak erre az eszközeik. Ez a problémakör visszacsatol a kiindulási pontomhoz, hiszen nagyon nehéz az A.I.-ok létrehozása során kikerülni az emberi hibákat.

Ráadásul az Agentic A.I.-ok működésének fontos aspektusa, hogy sokkal kisebb az ember ráhatása arra, amit az A.I. csinál. És míg egy rosszul sikerült kép egyszerűen újra elkészíthető - javítva, addig egy rossz döntés következményeivel más szinten kell szembenéznünk, a cselekedeteket nem tudjuk “csak úgy” törölni, vagy visszavonni.

A Mesterséges intelligencia “feltanítása” és működése

Az M.I. működésének röviden az az alapja, hogy egy adott, általunk (emberek) által ismert/létrehozott adatbázison “feltanítjuk” az M.I.-t, ami saját modelleket hoz létre, és ezen saját modellek alapján képes korábban nem ismert adatokat is létrehozni.

Azonban ebben rengeteg hibalehetőség rejlik. Csak néhány ezek közül:
- Rossz adatbázison történik a feltanítás (túl kevés adat, túl sok adat, hibás adatok, stb.).
- Ennek külön változata az, amikor nem mérjük fel, hogy az adatbázis mennyi emberi “gyarlóság”-ot, vagy rejtett érzelmet jelenít meg (így lehet például egy orvos-diagnosztikai M.I. rasszista). Az M.I. ezeket ugyanúgy beviszi a modelljébe, mint bármilyen más tényezőt, akkor is, ha a feltanítást végző személy ezt nem is tudatosítja magában.
- Lehet magának a feltanításnak a módszere is hibás, vagy nem optimális.
- Lehet az adatokhoz képest túl bonyolult modellre tanítani, vagy lehet, hogy a modell túlságosan jól illeszkedik az adatokra - ez a túltanulás. Ekkor az ismert adatokat majdnem 100%-ban visszadja az M.I., de új eredményeket nem képes generálni.
- Az M.I.-be beépített egyéb szabályok és eljárások nem illeszkednek a tanultakhoz.
- Direkt rosszul, vagy manipulatívan történik a feltanítás.
- Lehet rosszul feladatot adni egy M.I.-nek. A “rosszul” jelentheti, hogy nem praktikus, vagy az M.I. képességein/programozásán/belső szabályain túlmutató, vagy direkt romboló, esetleg manipulatív szándékú a feladat.

És így tovább, a sor eléggé hosszú. Amellett, hogy nagyon komoly kérdés, hogy az M.I.-k “nem megfelelő” működésének etikai és jogi felelősségének kérdésköre ma még szürke zóna - bár a jogi felelősség kapcsán az EU elég komoly lépéseket tett. És bizony, a felelősségek jó része a felhasználóra hárul.

Még egy dolgot szeretnék itt megemlíteni, mert akik használják az M.I.-t, azok vagy nem foglalkoznak azzal a kérdéssel, hogy nem megbízhatóak, vagy egyfajta vakhittel kezelik az M.I. által adott eredményeket. Pedig az M.I. működésének alapjai miatt az csak megadott százalékban fog helyes választ, vagy reakciót adni a kérésünkre. Amikor az M.I. téves választ/reakciót ad, azt hívják divatosan “hallucinálás”-nak.

Minden M.I. modell valamilyen valószínűségi működés alapját teremti meg. Ráadásul a feltanításnak is nagyon különböző változatai vannak, amelyek meghatározzák, hogy a modell hogyan “idomul” a feltanításhoz használt adatbázishoz, illetve magának az adatbázisnak a változásait is meghatározza. A három alapvető gépi tanulási módszer:
- Felügyelt tanulás
- Felügyelet nélküli tanulás
- (Félig felügyelt tanulás)
- Megerősítéses tanulás

A felhasználó számára ezen módszerek ismereténél talán fontosabb azt tudni, hogy az M.I. által adott válasz csak adott valószínűséggel jó.

Ezek a valószínűségek addig kielégítően magasak, ameddig nem távolodunk el a feltanításhoz használt adatbázis “hatóteréből” - matematikában ez az interpoláció.

Ha viszont elkezdünk eltávolodni ettől - matematikában ez az extrapoláció - akkor egyre rosszabb eredményeket kapunk. Hovatovább egy adott “távolságon túl” az eredményeknek a valószínűsége annyira lecsökken, hogy semmilyen formában nem elfogadható a válaszreakció. Nem véletlen, hogy a nagy chat M.I.-k folyamatosan tanulnak, folyamatosan bővítik az adatbázist, amire épülnek a modelljeik.

Érdekes például a chatGPT legfrissebb verziója körül kialakult felháborodás. Az új verziót egy megbízhatóbb, jobb, de “kisebb hatókörű” adatbázison tanították fel. Így a “standard” kérdésekben megbízhatóbb válaszokat adott, de nagyon hamar el lehetett jutni azokig a kérdésekig, amikre nem tudott helyes, vagy akárcsak értelmezhető választ adni. Lett is belőle botrány...

Mindezek nem misztikus dolgok, és alapszinten még csak nehéznek sem tekinthetőek (a konkrét matematikájuk azért időnként elég vad...), egyszerűen csak furák, és nem ismerjük őket eléggé. Például a nagy nyelvi modellek (Large Language Models - LLM) - mint például a chatGPT alapja is - a következő legvalószínűbb szót keresi. Ez nem bonyolult, de a következményeit csak tudatosan tudjuk feltárni,

önmagától, vagy “zsigerből” ez nem megy, nem olyan, mint egy kalapács, aminek gondolkodás nélkül is elég jól ismerjük a működését - és a nem megfelelő használat következményeit.

Zárszó

Mint általában a technikai eszközök esetén, az M.I. használatát sem utasítom el. Viszont fontosnak tartom, hogy ahogy megtanuljuk használni a többi eszközt (“kés, villa, olló, gyerek kezébe nem való...”), úgy az M.I. helyes használatát is el kell sajátítanunk. Annyi csak a nehézség, hogy ez esetben nincs gyerekkori tapasztalati hátterünk, de gyakran tudásunk sem, ami alapján a felelősségteljes használat kialakítható.

Mindezt nekünk kell tudatosan kialakítani, egyenértékűen a veszélyek (és kezelésük) feltárása kapcsán, és a sikeresélyes, jó használat kapcsán is (a kettő együttesen kezelendő).

És ehhez nem elegendő csupán a mesterséges intelligencia eszközök működését ismernünk, hanem azt is, hogy hogyan viszonyulnak hozzá mások, hogyan és mire használják. Ez is igaz egyébként minden egyes eszközünkre (kezdve egy kalapáccsal), de jellegéből fakadóan ez az A.I. tekintetében erősebben jelenlévő kérdéskör.

Hiszen használhatják manipulatív célokra ezeket, de valójában már a feltanítás is tartalmazhat félrevezető, manipulációs, vagy kifejezetten romboló célokat. És a legtöbb A.I.-al, vagy annak eredményeivel azonnal találkozhatunk az internetnek köszönhetően.

Nem szabad elfeledkezni az A.I. használatának társadalmi hatásaitól sem. Nagyon fontos a jelenkor fiataljainak tudni, hogy mely szakmákkal nem érdemes már foglalkozniuk (vagy helyette az adott szakma A.I. rendszereiről érdemes tanulniuk).

Ráadásul ebben nem csak az a kérdés, hogy az A.I. milyen munkák elvégzésére alkalmas, hanem az is, hogy mire fogják (vagy éppen nem fogják) a munkaadók használni (akár alkalmas rá, akár nem). Ergo, itt is megjelennek az eszköz működése és képességei mellett az emberi tényezők (akár egyénileg, akár társadalmi méretekben). És bizony, a közösségi média és a mobilkommunikáció megmutatta, hogy nagyon éretlenek vagyunk még ezeknek az eszközöknek a helyes használatára (tisztelet a nagyon ritka kivételeknek).

Láthatóan a téma annyira szerteágazó, és bizonyos tudományos/technológiai kérdések olyannyira kikerülhetetlenek, hogy ezt igazán működőképesen csak közösségben tudom elképzelni, ahol a különböző megközelítések összeadódnak, az ellentétes elgondolások ütköztetve vannak. Amíg ehhez partnereket nem találok, addig egymagam tévelygek tovább ebben a témában, amely meghatározza a közeljövőnket, és gyermekeink életét is.