Alapítva: 2001

INFORMÁCIÓS TÁRSADALOM

SZAKTUDOMÁNYOS FOLYÓIRAT

Juhos Sándor: Amikor a robot programozza az embert

Hivatkozás/reference:

Juhos Sándor, „Amikor a robot programozza az embert”, Információs Társadalom, XV. évf. (2015) 4. szám, 44-47. old.
http://dx.doi.org/10.22503/inftars.XV.2015.4.3

Reakció Z. Karvalics László a "Mesterséges intelligencia – a diskurzusok újratervezésének kora" című vitaindító írására.

When robots program humans

Comments to the paper by Z. Karvalics László, "Artificial intelligence – why to redesign the discourses?"

Juhos Sándor

Amikor a robot programozza az embert

A robotika iránti érdeklődésem kezdete még gyerekkoromra tehető. Akkor sem és most sem úgy gondolok a robotokra, mint független, öntudatra ébredő gépekre. Elképzelhető, ám ha be is következik, ehhez még el kell telnie legalább 10-15 évnek. Annak érdekében, hogy ez ne járjon negatív következményekkel, fontos, hogy megtegyük a megfelelő lépéseket.

Mitől is félünk?

Attól, hogy Asimov 3 törvénye kevés ahhoz, hogy megfelelő szabályozással kordában tudjuk tartani az önmódosító robotok működésének önállósodását. A robotok humanoidok vagy beágyazott rendszerek lehetnek, esetleg szoftverek, amelyekbe a programozók olyan genetikus algoritmusokat integrálnak, melyekkel képessé válnak az öntanulásra, vagy inkább önmódosításra. Mi a cél? Az, hogy óriási munkát igénylő programozással pontosan meghatározzuk a robot számára, mit tegyen, vagy az, hogy kevesebb munka után önálló cselekvésre ösztökéljük, s csak akkor módosítsunk a viselkedésén, ha az nem jó irányba halad?

Ha ez pusztán kényelem kérdése, akkor megteremtjük a saját problémánkat. Mert minél több dolgot hagyunk magától működni, annál kevesebb fölött marad meg a felügyeleti jogunk és lehetőségünk. Ha folyamatosan mi programozzuk a robotokat, akkor jobban beléjük tudjuk ültetni a logikát, a biztonságot, és könnyebben észrevesszük a lehetséges hibagócok kialakulását.

Öntanulás. Az öntanulás a robotokra nézve nem más, mint adatgyűjtés, amely a környezetből, a működésből, működtetésből fakad. Mivel nincs érzelmi tényező, csak igen/nem döntés a statisztikai eredmények által, a környezetből, és az adatbázisokból nyert információk összefüggései vezérlik és változtathatják meg a robotok működését.

Milyen szimulációkkal implementáljuk majd az algoritmusokat, hogy a véletlenszerű lehetőségek és variánsok mutálódását megállapítsuk? Mi történik akkor, ha a generálódás során nem sikerül a variánsok keletkezése során bekövetkező mutálódást leállítani – mert túl sok a variáció és a visszacsatolás. Követhetetlen az összefüggések logikája, mint a Pascal Triangle[1] esetében, egy végtelen folyamat indul be, amelyre nem számítottunk. Mi történik, ha a működés során bekövetkező változások meggátolják a hozzáférésünket?

[1] A Wikipédia a következőképpen fogalmazza meg a Pascal-háromszöget: „A háromszögben a sorok számozása zérótól kezdődik, és a páratlan és páros sorokban a számok el vannak csúsztatva egymáshoz képest. A háromszöget a következő egyszerű módon lehet megszerkeszteni: A nulladik sorba csak be kell írni az 1-est. A következő sorok szerkesztésénél a szabály a következő: az új számot úgy kapjuk meg, ha összeadjuk a felette balra és felette jobbra található két számot. Ha az összeg valamelyik tagja hiányzik (sor széle), akkor nullának kell tekinteni. Például az 1-es sor első száma 0 + 1 = 1, míg a 2-es sor középső száma 1 + 1 = 2”.

Az alap program megírása után a legtöbb öntanuló szoftver vizsgálja a környezetét. Képeket, tárgyakat, hangokat, összefüggéseket keres. Kell egy háttér adatbázis. De vajon mi lesz az? A világháló? Ha igen, akkor a már meglévő programokat, adattárakat, keresőoptimalizálásokat is használják majd? Vagy csak a képeket, a szavakat, a nyelveket, és ezek nyelvtani szabályait mint adattárat?

A legfőbb probléma, amelyet meg kell oldani, a hozzáférés. Erre nem azért van szükség, hogy a szoftver ne jusson hozzá bizonyos adatokhoz, hanem azért, hogy az illetéktelenek ne férjenek azokhoz hozzá a robot kommunikációs csatornáin és a szoftverén keresztül. Az Internet, mint lehetséges „agy", sokat segít az öntanulás és az érzelmi alapú döntéshozás szoftveres kifejlődésében. Igen ám, de ennek több a veszélye, mint az előnye. Az öntanuló szoftverek rengeteg anyagot, információt gyűjtenek, és a környezetükben keresik az összefüggéseket a működésük során. Statisztikákat vezetnek, prioritással rangsorolják a bejövő információkat – végül ebből születik egy eredmény, az eredményből pedig megkezdődik a végrehajtás.

Nézzük meg például a böngésző keresőmotorokat. Egy meghatározott algoritmus alapján megtanulják a szokásainkat, a kedvenc zenéinket, a kedvenc filmjeinket. Néha tévesen, mert az oldalak jellemzőit a Meta Title[2] és Meta Description Tag[3]-ek határozzák meg. Ha tehát a keresőoptimalizálás során több olyan kulcsszót adunk meg, ami nem 100%-osan a tartalomra utal, vagy nem teljesen jellemző az oldalra, téves statisztikát fog rólunk vezetni.

A legnagyobb probléma az, hogy a robotoknak nem tudunk érzelmen alapuló racionalitást tanítani. Még nem. Már a kezdetekben is kevésnek bizonyult Isaac Asimov 3 törvénye.

1. A robotnak nem szabad kárt okoznia emberi lényben, vagy tétlenül tűrnie, hogy emberi lény bármilyen kárt szenvedjen.
   (A jelentés mint "kár", erkölcsi kár is lehet, mert az erkölcs azonnal sérül, amint a robot felváltja a dolgozni akaró embert.)
2. A robot engedelmességgel tartozik az emberi lényeknek, kivéve, ha az utasítások az első törvény előírásaiba ütköznek.
3. A robot tartozik saját védelméről gondoskodni, amennyiben ez nem ütközik az első vagy második törvény bármelyikének előírásaiba.

   És a legrosszabb, ami történhetett, hogy megszületett a "Nulladik" törvény.
   „A robotnak nem szabad kárt okoznia emberi lényben, kivéve, ha valahogy belátja, hogy ez a kár végül az emberiség javára válik."

Hogyan látja be? Kiegészítésként az 1-2-3 törvényhez Asimov hozzátette, hogy nem oldhatják fel a 0. törvény tilalmát. Ki alkotja meg a programot, amely által egy robot belátja,

[2] Olyan nem látható címek, amelyek általában a weboldalra jellemző szavak és tartalmak az oldalon elrejtve, melyek a kereső optimalizálás során rávezetik a böngészőt, hogy a megadott kereső szavak alapján rátaláljon az oldalra.

[3] Olyan nem látható, a weboldal tulajdonságát tartalmazó leírások, melyekben tárolt adatok rávezetik a keresőmotort a találat kereséskor, hogy rátaláljon a weboldalra.

hogy mi és hogyan válik az emberiség javára úgy, hogy emellett káros egy egyén számára? Hogyan tervezi meg a robot a cselekedetet érzelem, képzelőerő nélkül, valamint meg nem történt állapotában a végkifejlet láncreakcióját és lehetséges következményét? A számozás azért lett a nulladik, mert magasabb sorszámú törvény nem írhat felül egy alacsonyabb sorszámút. Máris egy prioritás, amelyben a „nulladik" nem jó helyen van.

Az első törvény kimondja, hogy „*A robotnak nem szabad kárt okoznia emberi lényben*." A nulladik viszont belevisz egy olyan döntést, amit jelenleg még nem tudunk megoldani, de még azt sem tudjuk, hogy ezt a fajta döntéshozást milyen úton tanítjuk meg. Próbálkozunk feltárni az emberi agy működését, és megérteni az összefüggéseit, de azt elfelejtjük, hogy az ember évtizedekig fejlődik, és ezt követően mondható ki, hogy felnőtt. Képes az önálló életre, az önálló döntéshozásra. Amikor ez megtörténik, létrejön egyfajta ember, egyfajta viselkedésmód, egyfajta jellem.

Aztán Roger MacBride Allen átírta kicsit a 3 törvényt. Nem a nulladikkal kezdte, hanem kiegészítette egy negyedikkel.

1. A robotnak nem szabad kárt okoznia emberi lényben.
2. A robotnak együtt kell működnie az emberi lényekkel, kivéve, ha ez az együttműködés ütközik az első törvénnyel.
3. A robot tartozik saját védelméről gondoskodni, amennyiben ez az önvédelem nem ütközik az első törvénnyel.
4. A robot kedve szerint cselekedhet, kivéve, ha bármely cselekedete az első, a második vagy a harmadik törvényt sérti.

Azért a 4. törvény elgondolkodtató.

Amikor nem direkt, csak az ember szolgálatában áll, azt csinálhat, amit akar. Illetve az ember igénye szerint néha irányított, néha szabadon tevékenykedhet. Ez sem teljesen jó lehetőség. Gondoljunk bele, mi lenne, ha az egyik robot inputja az összes többi robot outputja. Így egymás tapasztalatait is begyűjtenék. A folyamat követhetetlenné válna.

Kezdjük az elején!

Az ember. Mivel viszonylag jól működünk, jó példa lehetünk a robotok számára. Igaz, sok olyan dolog van, amit a robotok még jó ideig nem lesznek képesek kivitelezni az ember viselkedéstanulásából. Megszületik, növekedése során gyűjti a környezetéből az információkat. A születés egyenlő a robot létrehozásával.

A szülő folyamatosan követi gyermeke életpályáját. Segítik a családtagok, az iskola, a törvényhozás, az államigazgatás. Az erkölcstan. Tehát ha a gyerek tapasztal valamit, és megkérdezik, hogy a tapasztalásból mit ért meg, mit szűr le, és ezt követően hogyan cselekszik, befolyásolható, és a helyes útra terelhető, ha tévesen értelmezett egy tapasztalást. Igen ám, de ki határozza meg, hogy mi a helyes? Mint ahogy a szülő hibáját is orvosolja az iskola, a programozók téves logikai összefüggésrendszerét is ellenőriznie kell egy robotikai felügyelő szervnek, mert itt már kevés a három (négy) törvény.

A helyes irány az lenne, ha minden egyes önmódosító folyamatot csak az ellenőrzés és vizsgálat után – nem okoz-e kárt – engedélyeznénk. Nem csak műszaki, fizikai kárra,

vagy veszélyre gondolok, hanem azt is szűrni kell, hogy a "master"[4] mindig az ember, a "slave"[5] mindig a robot maradjon. Tehát a lényeg, hogy a robot nem kelhet önálló "életre", és nem végezhet kontrollálhatatlan, kiszámíthatatlan cselekvéseket. A gond csak az, hogy mint minden rendszer a világon, megkerülhető, kiiktatható, és az érdekek iránymutatása szerint befolyásolható.

Hogyan állítsunk fel egy olyan felügyelő rendszert, amelyen nincs olyan kiskapu, melyen keresztül a szoftver megkerüli a felügyeletet? A legjobb példa erre a vírus, és ellenszere, a vírusirtó. De a kulcs nem a vírusban és a vírusirtóban keresendő. Tehát ne orvosoljuk a hibát, mint lehetséges megoldást, inkább védje az alapszoftvert egyfajta tűzfal a tanult változásoktól, és csak akkor engedje beépülni a változást, ha auditálva lett. Tehát olyan robot tűzfal kell, amely észreveszi a negatív kimenetelű változást.

Vagy hogy jobb hasonlattal éljek, említhetem az automatikusan aktiválódó szoftverfrissítést. Ha nem akarom, hogy automatikus legyen, értesítést és részletes jelentést kérek minden változási szándékról. Egy öntanuló robotnál azt lehetne tenni, hogy az öntanulási statisztika által végbemenő változások egy tárba kerüljenek, és csak akkor aktiválódjanak, ha a felügyeleti szerv által biztonságosnak lett ítélve.

Vannak olyan agyi folyamatok, cselekedetek, melyek életünk során csak ritkán vagy soha nem következnek be. Ilyen például egy másik ember életének kioltása. Generációs nevelés folyamatának eredménye, hogy „jó embert alkossunk." A jó robothoz is sok-sok idő kell. Jó programozó, jó szándék és jó cél. A robotot, és annak felhasználását annak idején Karel Čapek sem így képzelte el. A rossz útra akkor kerültünk, amikor egy robot nem egy embert helyettesített, hanem több(et). A jó felállás az lenne, ha minden ember egy tanár volna, és minden robot egy diák, az ember „manipulátora".

Vegyük például a vadállatokat vagy a már megszelídített, háziasított állatokat. Elindult egy nevelési folyamat, állandó felügyelettel, melynek folyamán háziasítottuk, kineveltük az állatokból az agresszivitást, az ember számára negatív tulajdonságokat. Az evolúció folyamán a rosszat elsorvasztotta, elnyomta a nevelés, a folyamatos kontroll, és a vadállat folyamatos figyelmeztetése. A kutya szelídített állat, ám az ősgénekbe kódolt vadállati viselkedés a mutáció során visszaváltozhat, és ha nincs meg a kellő kontroll, az állat ismét kiszámíthatatlan vadállattá lesz.

Tisztában vagyunk azzal, hogy a háziállatok hogyan viselkedtek megszelídítésük előtt, ennél fogva óvatosak maradunk velük, ott van bennünk a félsz, egy veszélyérzet, amely arra figyelmeztet bennünket, hogy éberek maradjunk. A tisztes távolságot az öntanuló robotokkal is meg kell tartanunk, és megfelelő mennyiségű biztonsági intézkedéseket kell tennünk velük kapcsolatban.

Juhos Sándor 1975-ben született Győrben. A győri Technics Playground Robotikai Automatizálási és Mechatronikai Oktatóközpont vezetője, a Robotika Szakosztály alelnöke. Az általa épített humanoiddal csapatával a 2009-es RoboCup világbajnokságon első helyezést értek el, SuperTeam világbajnokok lettek.

[4] Mester, tanító, akinek a tudásából fakadó kialakított hierarchia szerint mindig engedelmeskednek.
[5] Szolga, aki a mesternek mindig alárendeli magát.