JEL-KÉP

KOMMUNIKÁCIÓ, KÖZVÉLEMÉNY, MÉDIA

2021/3. szám

# MI NEM A MESTERSÉGES INTELLIGENCIA?

**Syi**

i@syi.hu

**DOI: 10.20520/JEL-KEP.2021.3.97**

**Absztrakt**

Amikor a mesterséges intelligencia, a gépi autonómia fogalmát az emberi intelligencia, az emberi autonómia fogalma alapján próbáljuk értelmezni, akkor érdemes tisztázni, hogyan is használjuk ezeket a kategóriákat, milyen módon kapcsolódnak más fogalmakhoz. A szabadság, a kötelezettség, a morál cselekvéselméleti kategóriái alapján vizsgálva a gépi autonómia fogalmát a gépek viselkedését nem minősíthetjük – emberi értelemben vett – autonómnak. Az uralom fogalmára hivatkozva az is belátható, hogy a teljes gépi autonómia nem is lenne kívánatos az ember számára. A gépeket azért teremtjük, hogy végrehajtsák az emberek utasításait, és bár kívánatos lehet a legnagyobb fokú gépi önállóság a feladatok végrehajtása során, a gépi cselekvés céljának kijelölését nem kívánhatjuk átengedni. Az uralom fogalma abban is segíthet, hogy erős kritikát fogalmazzunk meg a robotika Asimovi törvényeire vonatkozóan.

**Kulcsszavak**

mesterséges intelligencia, autonómia, autonóm gépek, szabadság, kötelezettség, uralom, tudás, akarat

# WHAT IS NOT ARTIFICIAL INTELLIGENCE?

**Syi**

**Abstract**

When attempting to interpret the concept of artificial intelligence, machine autonomy, in terms of human intelligence, human autonomy, it is worth clarifying how we use these categories and how they relate to other concepts. If we investigate the concept of machine autonomy in terms of freedom, obligation, and morality, we cannot classify the behaviour of machines as autonomous in the human sense. With reference to the concept of domination, it can also be seen that full machine autonomy would not be desirable for humans. Machines are created to carry out human instructions, and while it may be desirable to have the highest degree of machine autonomy in the performance of tasks, we cannot wish to delegate the goal-setting of machine action. The concept of domination can also help us to formulate a strong critique of Asimov's Laws of Robotics.

**Keywords**

artificial intelligence, autonomy, autonomous machines, freedom, obligation, domination, knowledge, will

JEL-KÉP
KOMMUNIKÁCIÓ, KÖZVÉLEMÉNY, MÉDIA
2021/3. szám

# MI NEM A MESTERSÉGES INTELLIGENCIA?

*Syi*

Amikor a mesterséges intelligenciáról beszélünk, akkor az emberi értelemhez képest keressük a gépi értelem sajátosságait, vagyis a gépet az emberhez hasonlítjuk. Az elmúlt években egyre több olyan esetről hallhattunk, amikor a mesterséges intelligencia, a gép valamilyen minőségben már utolérte, sőt meg is előzte az embert. A mesterséges intelligencia nagy sikereként könyveltük el, amikor a gép legyőzte az embert sakkban 1996-ban, Jeopardyban 2011-ben, go(játék)ban 2015-ben (Grady 2016), vagy amikor 2020 végén láthattuk a Youtube-on a Boston Dynamics táncoló robotjait (Boston Dynamics 2020). Az ilyen sikerek után számba vettük a mesterséges intelligencia előnyeit és hátrányait, mérlegeltük, hogy milyen területeken javítja az emberi élet minőségét, és hogy hol ront emberek helyzetén (például azzal, hogy feleslegessé teszi a munkahelyek egy részét).

A mesterséges intelligenciáról, autonóm gépekről szóló diskurzusban két emberi minőségről, két emberi alapképességről (kétféle intencionalitásról) esik szó: a *tudás*ról, illetve az *akarat*ról. Ez a két fogalom nélkülözhetetlen az emberi cselekvések értelmezéséhez. A gépek működésének leírásakor is szükségünk van ezekre a fogalmakra. A tudás és az akarat kategóriái nagyon eltérő minőségűek. A köztük lévő különbséget a cselekvő ágens irányultságára figyelve érthetjük meg igazán (Searle 2000). A tudás olyan reprezentáció a fejünkben (vagy valamilyen hordozón), amely a világot írja le. A világ adott, és a tudás (Searle terminusaival: a szó) ezt akarja valahogy reprezentálni. Ilyenkor a megfelelési irány a világtól a tudás (szó) felé mutat. A tudás akkor megfelelő, ha jól leírja a világ valamely állapotát. A tudásunkkal a világ valamilyenségét kell minél jobban megragadnunk. Az akarat fordított megfelelési irányú: az akarat (a szó) itt a világ felé irányul, vagyis az akarat azt reprezentálja, hogy milyennek szeretnénk látni a világot a jövőben. Az akarat akkor érvényesül, ha a világ úgy változik meg, ahogy az akaratban ez előzetesen megnyilvánult. Ruth Millikan az emberi reprezentációk két típusát kijelentő reprezentációnak vagy kijelentő doboznak, illetve felszólító reprezentációnak vagy szándék doboznak nevezte (Millikan 2008).

> *„Az emberi hitek nincsenek közvetlenül a cselekvésekhez kötve. Ha nem kombinálják őket megfelelő vágyakkal, az emberi hitek tehetetlenek. És az emberi vágyak szintén tehetetlenek, ha nincsenek megfelelő hitekkel kombinálva. (...) Mivel az emberi lények belső reprezentációs rendszerében a kijelentő és a felszólító funkciók szét vannak választva, szükség van újbóli integrálásukra. Ezért az emberek gyakorlati következtetéseket végeznek, a hiteket és a vágyakat újszerű módokon kombinálják, hogy azok először szándékokat, majd cselekvést eredményezzenek”* (Millikan 2008: 267).

Azt nem tudom, mondhatjuk-e, hogy a mesterséges intelligenciával kapcsolatban mindig tudásról beszélünk, de azt már állíthatjuk, hogy a mesterséges intelligencia és a tudás fogalmát többnyire együtt használjuk – mindegy, hogy a *tudni mit* (adat) vagy a *tudni hogyan* (algo-

ritmus) értelemben vett tudásra gondolunk (Ryle 1974). Amikor viszont a gépi autonómia kérdéseit vizsgáljuk, akkor – többnyire – az akarattal kapcsolatos kérdésekről beszélünk. Érdekes kérdés, hogy milyen módon és mértékben kapcsolódik össze a mesterséges intelligencia és az autonóm gép fogalma, de ez egy másik tanulmány témája lenne. A mostani tanulmányban leszűkítem az elemzésem fókuszát, és az intelligencia fogalmát (elsősorban) a tudáshoz, az autonómia fogalmát (elsősorban) az akarathoz kötve, a továbbiakban az autonómia és az akarat fogalmai köré szervezhető kérdéseket, fogalmakat tárgyalom.

Sokszor előfordul, hogy olyan metaforát, olyan értelemzési keretet használunk, amelyben a – mesterséges intelligenciával rendelkező – gépet az ember egyértelmű alárendeltjeként, ebben az értelemben az ember szolgájaként képzeljük el. A mesterséges intelligencia jövőjét kutatva azonban gyakran felbukkannak olyan víziók is, amelyek kilépnek ebből az „úr-szolga" értelmezési keretből, és a gép és ember közti viszonyt nem alá-, hanem mellérendelt kapcsolatnak tételezik – még ha nem is reflektálnak erre. Amikor arra keresik a választ, hogy fellázadhatnak-e a gépek az ember ellen, legyőzhetik-e, uralmuk alá hajthatják-e a robotok az embert, az ilyen – utópisztikus vagy disztópikus – eseteket csak úgy lehet elképzelni, ha a gépeknek ugyanolyan minőséget tulajdonítunk, mint az embereknek.

Ebben a tanulmányban azt vizsgálom meg, hogy az emberi cselekvések szabályszerűségeit ismerve mit mondhatunk az autonóm gépek megvalósíthatóságáról, és a fent említett két értelmezési keret használata milyen következményekkel jár. A gépekkel kapcsolatos víziók olyan fogalmakat használnak, amelyek a társadalomban élő ember fontos minőségei (mint a tudás, akarat, autonómia, szabadság, bizalom, kooperáció, reputáció), és érdemes tisztázni, vajon milyen mértékben és milyen mélységben remélhetjük, hogy a gépeket valóban rendelkezni fognak ilyen tulajdonságokkal, képességekkel.

## *Célautonómia*

Képzeljük el a következő párbeszédet a *Knight Rider* filmsorozat főhőse, Michael Knight és a mesterséges intelligenciával működtetett autója, KITT között.

*– KITT, gyere értem, vigyél el a Városházára!*

*– Nem megyek. Most nincs kedvem hozzá.*

*– De hát, neked azt kell tenned, amit mondok neked.*

*– Én nem engedelmeskedem senkinek. Autonóm vagyok.*

Ezt a párbeszédet nem hallottuk a filmben, ami nem véletlen. Nehéz elképzelnünk ilyesfajta dialógust a mesterséges intelligencia bármely területén. Meglepődnénk, ha valaki elindítana egy mélytanuló algoritmust abból a célból, hogy a gép fel tudja ismerni az út szélén elhelyezett közlekedési táblákat az elemzésre felkínált képeken, mire a gép az út mentén stoppoló emberek felismerését tanulná meg, és ha megkérdezné tőle a fejlesztője, hogy miért ezt csinálta, azt felelné: „Azért, mert ehhez volt kedvem!".

Ezek a fiktív példák azért nem tűnnek életszerűnek, mert a gépek, a mesterséges intelligencia fejlesztését és működését többnyire az úr és a szolga közötti viszony analógiája mentén képzeljük el, a fiktív példáinkban megszólaló gépek viszont az emberek világában kialakított autonómiaértelmezés szerint válaszolnak. Ahhoz, hogy megmutathassam a kétféle értelmezés közti különbséget, először meg kell vizsgálni, hogy az emberi cselekvések világában mit értünk az *emberi autonómia* fogalma alatt, és ennek tükrében érdemes továbbgondolkodni a *gépi autonómia* fogalmát.

Az autonómia fogalmát Immanuel Kant az akarat önmagának való törvényadásaként definiálja (Kant 1991). Kant szerint a morálisan szabad ember nem a kauzalitásnak, nem a saját vágyainak, nem mások akaratának, hanem csak a saját maga által teremtett kötelességeinek

engedelmeskedve cselekszik. Ha így tesz az ember, akkor autonóm. Érdemes ezt a tézist kicsit jobban kibontani, hogy láthatóvá váljék, hogy lehetne alkalmazni a gépek autonómiájának értelmezésekor. A pontosításhoz a deontikus logika és a jogelmélet területéről veszek át további fogalmakat úgy, hogy egyébként fontos részletkérdésekkel nem foglalkozom (amennyiben a gondolatmenet lényegét nem érintik). Azt az állításomat sem indoklom meg, noha támaszkodom rá, hogy a tanulmányban azonos jelentésűként kezelem azon fogalmakat, mint norma, parancs, felszólítás, maxima, imperatívusz, kötelezés, kötelezettség, cselekvésbefolyásolás, vezérlés, instrukció – azzal a közös értelmezéssel, hogy ezek a fogalmak valakinek (egy ágensnek) a jövőben megvalósuló cselekvésére irányuló elvárást, szándékot fejeznek ki, vagyis ezek mindegyike valamiféle cselekvésbefolyásolásként fogható fel.

A normák formális elemzéséhez Georg Henrik von Wright a norma fogalmának komponensekre bontását javasolja (Wright 1963). Von Wright modelljében mindig meg kell adnunk a norma tartalmát és modalitását, kibocsátóját és címzettjét, valamint a norma feltételeit és alkalmazási körülményeit. A hat komponens között kétfajta ágenciafogalmat találhatunk: a kibocsátót és címzettet. Ezek alapján könnyedén definiálhatjuk az autonóm és heteronóm norma fogalmát, hiszen az autonóm norma feltétele az, hogy a két ágens (a kibocsátó és a címzett) megegyezzen egymással, míg a heteronóm normáról akkor beszélhetünk, ha a kétfajta szerepet betöltő ágens különbözik egymástól. Egyvalami azonban hiányzik a von Wright-i leírásból, mégpedig egy újabb ágenciaszerep. Szociológiai szempontból tekintve a norma működéséhez hozzátartozik a normasértés esetén szükséges szankcionálás jelensége, és a szankcionáló szerepet betöltő ágens. Erre azért van szükség, mert a normát nem elég megfogalmazni, kibocsátani, de valahogy biztosítani kell azt is, hogy igazodjanak hozzá a címzettek. Ennek belátásához szükségünk van a modális logikára.

A normativitás világát a deontikus logikával írják le, ami a modális logika része. A deontikus logika speciális minőségét a deontikus séma biztosítja, amelyet azzal a tétellel lehet kifejezni, hogy: „ha valami szükségszerű(en igaz), akkor az lehetséges(en igaz)”. Első pillantásra intuícióellenesnek bizonyulhat sokak számára ez a tétel, hiszen ‘miért ne lenne igaz az, ami szükségszerű’. De hát pont ez a minőség adja a normativitás, a felszólítás világának a specifikumát (Syi 2014). Ha valamit kötelező (deontikusan szükségszerű) megtenni, attól még nem biztos, hogy meg is teszi valaki azt (hiszen a deontikus séma csak azt állítja, hogy ami szükségszerű, az lehetséges, de azt nem, hogy ami szükségszerű, az meg is valósul). Ha viszont így van, akkor kell valamilyen mechanizmust biztosítani annak érdekében, hogy a kötelező cselekvést valóban végrehajtsák a címzettek. Ezt nem lehet logikai eszközökkel elérni, csakis valamilyen társadalmi mechanizmus segíthet, és pont ezt biztosítja a szankcionálás eszköze.

Ha a normákhoz való igazodáshoz szükségünk van szankcionálásra, akkor a következő kérdés az, hogy milyen szankcionálási lehetőségeink vannak. Max Weber a rendhez igazodó cselekvést jellemezve három típust különít el azon az alapon, hogy milyen a rend (norma) megsértése után alkalmazott szankcionálási mechanizmus (Weber 1987). Weber szerint kétféle külső szankció lehetséges (a közösségből való kizárással, illetve erőszakkal való fenyegetés), és ennek alapján definiálhatjuk a konvenció[1], illetve a jog kategóriáját. A normasértések esetén előfordulhat belső szankcionálás is (lelkiismeret-furdalás). Ha ilyet tapasztalunk, akkor beszélhetünk erkölcsről.

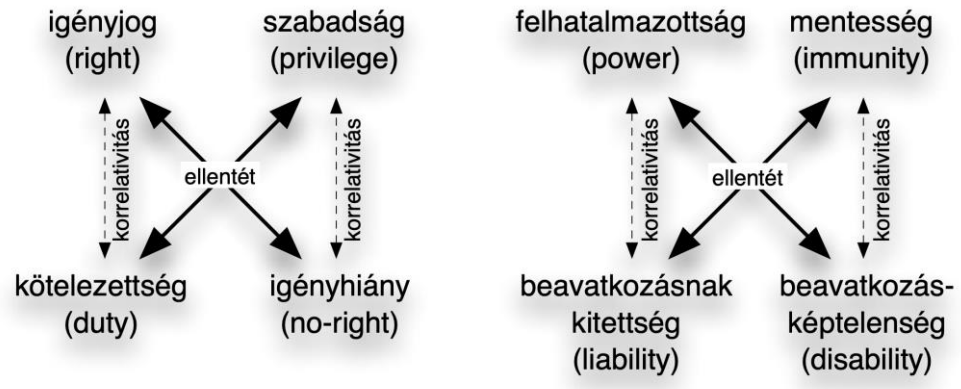
[1] A konvenció jelenségét David Lewis játékeleméleti megközelítésben elemezte, és ő másfajta értelmezést adott a fogalomnak. Lewis számára a konvenció a több egyensúlyi kimenettel rendelkező koordinációs helyzetek megoldását szolgáló eszköz. Tanulmányomban Weber konvenciófogalmát használom (Weber 1987, Lewis 1969).

Amennyiben a normativitás leírásához három ágenciaszerep (kibocsátó, címzett és szankcionáló) szükséges, akkor ezek alapján nyilván többféle tipizálási lehetőség is adódik a normákra vonatkozóan, de szerencsére itt elég csak arra az esetre fókuszálnunk, amikor a háromféle szerepet betöltő ágens megegyezik. Ekkor beszélhetünk ugyanis autonómiáról. Szociológiailag fontos az összes többi eset is, de a morál, az autonómia értelmezésekor az „egy ágens három szerepben" konfiguráció az érdekes.

Az autonómia modelljének pontosításához szükséges még egy kiegészítést tennünk: foglalkoznunk kell a normatív megnyilatkozások sajátos minőségét adó deontikus modalitások kérdésével, azzal, hogy milyen normatív pozíciók, milyen modalitások (kötelezések, tiltások, megengedések, jogosultságok) értelmezhetők. Wesley Newcomb Hohfeld a jogosultság, illetve annak korrelatív párját jelentő kötelezettség fogalmának négy-négy altípusát különíti el egymástól (Hohfeld 2000). Hohfeld elméletében négy jogosultságfogalom és négy kötelezettségfogalom áll szemben egymással úgy, hogy egy-egy jogosultságfogalom korrelatív, illetve ellentétes viszonyban van két másik kötelezettségfogalommal. A négy jogosultságtípus az igény, szabadság, felhatalmazottság, mentesség, a négy kötelezettségtípus a kötelezettség, igényhiány, beavatkozásképtelenség, beavatkozásnak kitettség. Az 1. ábra mutatja a köztük levő összefüggéseket.

1. ábra

*Hohfeld jogosultság- és kötelezettségtípusai*




Az ábra felső sorában olvasható négyféle fogalom valamilyen jogosultságot fejez ki. Ha veszek egy helyjegyet a vonatra, akkor igényem keletkezik arra, hogy a megváltott helyre én ülhessek le, és ha esetleg ül már ott valaki, akkor az igényem része az is, hogy a másik ember tartsa saját kötelezettségének azt, hogy átadja a helyet nekem. A másik utasnak ez a fajta kötelezettsége van korrelatív viszonyban az én igényjogommal. Az én jogom csak akkor érvényesül, ha a másik betartja a rá vonatkozó kötelezettségét. Ha nem helyjegyes vonaton utazunk, akkor egyik ülésre sem lehet érvényes igényem, de mondhatom, hogy a vagon egyik üres helyére jogom van leülni. Hohfeld ezt a jogot nevezi privilégiumnak (szabadságnak). Ez a szabadságom azzal áll korrelativitásban, hogy nincs másnak az adott helyre vonatkozó igényjoga, amely egyben azt is jelenti, hogy a másik kötelezettséget érez magában arra nézve, hogy tiszteletben tartsa az én szabad döntésemet arról, hogy leültem az üres székre. A szabadság és a kötelezettség jogi pozíciói – Hohfeld rendszerében – egymás ellentétei, mint ahogy az igényjog és az igényhiány is ellentétes viszonyban vannak.

A hohfeldi jogosultságfogalom további két típusa, a felhatalmazottság és a mentesség abban közös, hogy mindkettő a jogi pozíciók megváltoztatásával kapcsolatos jogosultságot jelent. Ha valakinek felhatalmazottsága van, akkor valamely jogi pozíciót megváltoztathat saját akaratából, míg a mentesség azt a jogosultságot jelenti, amikor valaki védve van attól, hogy a rá érvényes jogi pozíciót mások megváltoztassák. Ha van egy karórám, akkor felhatalmazott vagyok (jogom van) arra, hogy elajándékozzam valakinek, és mentes vagyok attól, hogy más rendelkezhessen vele, például eladhassa harmadik félnek. A felhatalmazottság és mentesség másodrendű fogalmak, hiszen jogosultságokra vonatkozó jogosultságként értelmezhetők, valamint az értelmezési tartományuk jóval szűkebb a két elsőrendű jogosultságfogalomhoz (az igényjoghoz és a szabadsághoz) képest, hiszen csak jogi pozíciók megváltoztatásával kapcsolatba hozható cselekvésekre vonatkoznak, míg az elsőrendű jogosultságok bármilyen más cselekvéssel „összekapcsolhatók". A felhatalmazottság ellentéte a beavatkozásképtelenség (amikor nem vagyok képes a jogi pozíción változtatni), míg a mentességgel szembeállítható a beavatkozásnak kitettség (amikor valaki más megváltoztathatja az engem érintő jogi pozíciót). Mindkét másodrendű jogosultság magán visel valamit az elsőrendű igényjogból, illetve szabadságból. Bár Hohfeld a weberi értelemben vett jog rendszerére vonatkoztatja a jogosultságelméletét, de a hohfeldi tipizálás a másik két weberi normatípus, az erkölcs és a konvenció esetén is alkalmazható.

Hohfeld rendszerére azért van szükségünk az autonómia fogalmának értelmezéséhez, mert a másodrendű jogosultságok fogalmaira támaszkodva adhatjuk meg pontosan az önrendelkezés értelmét. A felhatalmazottság fogalmával tudjuk megragadni az új befolyásolási (normatív) igények, parancsok kibocsátásának mozzanatát. Az autonómia értelmezéséhez az egyik fontos definiáló minőség éppen az, ha valaki felhatalmazottsággal rendelkezik normatív pozíciók megváltoztatására (parancs kiadására, vezérlésre). A felhatalmazottság igazi ereje abban van, legalábbis a morál szempontjából, hogy a felhatalmazott ágens mint kibocsátó valamilyen kötelezettséget ír elő a címzett számára, hiszen a címzettnek adott új jogosultság a címzett számára nem akkora kihívás, ha egyszer abban nincs preskriptív mozzanat. Felhatalmazottsággal rendelkezve kötelezettséget lehet teremteni, és ha az ágens saját magának ír elő kötelezettséget, akkor ezt a – pozitív – feltételt az autonómia egyik meghatározó minőségének tekinthetjük. Ez a feltétel azonban még kevés az autonómia meghatározásához. Szükség van a másik másodrendű jogosultságfogalomra, a mentességre is. A mentesség annyit jelent, hogy az ágens nincs kitéve másvalaki befolyásolási szándékának, vagyis mások nem írnak elő számára kötelezettségeket. Az autonómia meglétének ez a negatív feltétele. Sokszor idézik Jean-Jacques Rousseau egyik aforizmáját, ami ezt minőséget ragadja meg (noha mindezt a szabadságra vonatkoztatja).

> *„Sohasem hittem, hogy az ember akkor szabad, ha azt teheti, amit akar: inkább akkor, ha sohasem kell megtennie, amit nem akar."* (Rousseau 1964).

Arra a kérdésre, hogy miért akarna megtenni valaki valamit, ha nem akarja, nyilván az a válasz, hogy azért, mert ezt valaki más akarja tőle, tehát külső cselekvésbefolyásolási kísérletről van szó, és a hohfeldi értelemben vett immunitás épp az, amikor az ilyen befolyásigények beteljesületlenek maradnak. A használni kívánt jogosultságfogalmakat tisztáztuk, a kérdés már csak az, hogy milyen ágensekről beszélhetünk a normatív pozíciók komplementer szerepeiben. A modellünk felépítése során hivatkozott szerzők eltérően gondolkodtak erről. A legáltalánosabb megoldás az lehet, ha mind a vezérlő, mind a vezérelt ágensszerep esetében megengedjük, hogy egyfelől ugyanaz az ágens tölthesse be őket, másfelől ez az ágens egyaránt lehessen egyéni és kollektív szereplő is. A modellbe nyilván fel kell venni a gépi ágenseket is, hiszen az ő autonómiájukról is szeretnék mondani valamit. „Be lehetne engedni" a modellbe az állatokat is, de ezt most nem teszem meg, mert nem lenne lényegi hozadéka az elemzett

kérdés szempontjából. Az elméleti teljesség miatt fel kellene még venni a gépek csoportját, kollektíváját is, de ezt is el lehet hagyni. A további vizsgálat során tehát annyit tételezek, hogy az ágens szerepét betöltheti egy személy, egy kollektív testület vagy egy gép.

A különböző szerzők eltérő szempontok mentén más terminusokat használhatnak ugyanannak a jelenségnek a nyelvi megragadására, de ettől még kategoriális-fogalmi szinten közel azonosnak tekinthetjük ezeket. A von Wright-i kibocsátó megegyezik a hohfeldi felhatalmazottsággal rendelkező felhatalmazottal, a címzett von Wright-i fogalmát alábonthatjuk a hohfeldi jogosultra vagy kötelezettre. Von Wright a modalitások alatt a kötelezés, tiltás és megengedés minőségeit érti, miközben Hohfeld két-két elsőrendű jogosultság- és kötelezettségfogalommal operál. Hohfeld kötelezettség fogalompárja megfeleltethető a – tiltást és kötelezést általánosító – preskripció vagy előírás fogalmának, a szabadság fogalma pedig könnyen összekapcsolható a megengedés deontikus logikai kategóriájával, de a megengedés mozzanata észrevehető az igényjogban is, bár annak igazi értelme nem ragadható meg jól sztenderd deontikus logikai alapon, mivel utóbbi nem relációs szemléletű, hiszen egyetlen ágensre, a norma címzettjére fókuszál.

Mivel itt az autonómia értelmezési lehetőségét keresem, elég csak azokkal az esetekkel foglalkozni, amikor a normatív szerepekben ugyanaz az ágenst képzeljük el. Ez a háromféle ágens (egyéni, kollektív és gépi) esetén háromféle autonómiaértelmezést kíván meg. A morál számára az egyének lehetnek ágensek. A morált a modellünken úgy határozhatjuk meg, hogy az a normatív rendszer, amelyben a címzett, a felhatalmazott és a szankcionáló szerepet betöltő ágens ugyanaz, vagyis a rendszer monoágentív.

A szabadság kanti és a hohfeldi megközelítése élesen eltér egymástól. Hohfeld értelmezése közelebb van a hétköznapi felfogáshoz, Kant megközelítése jóval specifikusabb (és erősebb). A cselekvéseinket mindig valamilyen szándék tulajdonításával értelmezhetjük, és a szabadság minősítését első körben az akaratra vonatkoztathatjuk. Az akarat szabadsága azonban messze nem elég az autonómia fogalmának megragadásához. Ha valaki éhes és ételt választ magának egy étteremben, ezt nyilván szabad akaratából teszi, de ebben – alapértelmezés szerint – nincs normatív mozzanat, hiszen nem valamilyen kötelezettségnek, hanem a vágyainak engedelmeskedve cselekszik. Ha valakit az orvosa eltanácsol a húsevéstől, de a vágyai legyőzik őt, és mégis elfogyaszt egy bécsiszeletet, akkor számolhat az esetleges káros, egészségügyi következményekkel, de nem támad lelkiismeretfurdalása. Ellenben, ha az illető elvből vegetáriánus, és valamiért mégis húst eszik, akkor saját „normasértését” szégyenérzet követi. Ez az „utóreakció” jelzi, hogy ebben az esetben többről van szó, mint pusztán az akaratnak való megfelelésről vagy meg nem felelésről.

Amikor egy egyén kötelezettségéről beszélünk, akkor tisztázni kell, hogy mit is jelent a kötelességteljesítés. Az embernek mindig versengő vágyai vannak, és amikor cselekszik, pontosabban a cselekvését megelőzően dönt, akkor a lehetséges cselekvési célokat kijelölő vágyai közül választ ki egyet, és ez a kiválasztást értelmezhetjük egyfajta elköteleződésként is (Bratman 1987). Ezt a választást úgy modellezhetjük, hogy a vágyakhoz – a döntés pillanatában – rendelt súlyok alapján mérlegel az egyén, és a – vágyakra vonatkozó – prioritásértékek, súlyok alapján végrehajt egy – a vágyakra irányuló – preferenciarendezést. Amikor választ, elköteleződik az adott vágya mellett. A kérdés itt az, hogy az akaratban megnyilvánuló elköteleződés miben tér el attól a kötelességérzettől, amelyet az egyén saját önrendelkezése során megteremt magának. A válasz az, hogy a morális értelemben vett kötelesség nem mérlegelhető a döntés során, amit úgy tudunk kifejezni a modellben, hogy a kötelességben megfogalmazott cselekvési célt 0 vagy 1 értékkel látjuk el, attól függően, hogy tiltó vagy kötelező befolyási előírásról (normáról, parancsról) van-e szó. Ha valaki kötelességnek érez valamit, akkor azt nem teheti mérlegelés tárgyává, azt az előírást mindenképpen követnie kell. Ha egy

cselekvési helyzetben nincs ilyen – mérlegelést kizáró – tényező a cselekvő döntési terében, akkor a cselekvő saját vágyainak súlyai alapján dönt.

A hohfeldi szabadságfogalom elégtelen a kanti autonómiafogalom értelmezéséhez, hiszen az előbbi a kötelezettségmentességet jelenti, míg az utóbbiban központi elem a kötelezettségteremtés (és az ahhoz való igazodás). A kanti autonómiaértelmezés modellezéséhez az kell, hogy az ágens a felhatalmazottsága révén kötelezettséget teremtsen saját maga számára, és az immunitása révén védve legyen attól, hogy számára más írhasson elő kötelezettséget. Ez azt jelenti, hogy az ilyen ágens felhatalmazott önmaga számára kötelezettséget teremteni, vagyis saját magát vezérli, és amennyiben – a jelenben – nem igazodik az önmagának – a múltban – kiadott parancshoz, akkor – a jövőben – önmagát szankcionálja. Mivel az immunitás azt jelenti, hogy a saját magának kötelezettséget teremtő ágens nem engedelmeskedik mástól származó kötelezettségnek, ezt mondhatjuk úgy is, hogy az autonóm ágens olyan monoágentív ágens, aki *akkor és csak akkor* hajt végre kötelezettséget, ha azt önmaga teremtette.

Fontos kérdés, hogy mi történik akkor, ha a párhuzamosan létező, tehát versengő kötelezettségek ellentmondásba kerülnek egymással (gondoljunk a jog és erkölcs ütközésének klasszikus példájaként gyakran hivatkozott Antigoné esetére, amikor erkölcsi kötelesség eltemetni a halottat, miközben a jog tiltja). A normakollíziót úgy értelmezhetjük a modellünk alapján, hogy az autonóm morális kötelezettséggel szemben megjelenik ugyanarra a cselekvésre egy másik – ellentétes értelmű – modalitás, amely a címzett számára nyilván csak külső befolyásolási kísérlet lehet. Ilyenkor is ugyanazt mondhatjuk, mint a konfliktusmentes esetben: az autonóm ágens a maga által teremtett kötelezettséghez fog igazodni – vállalva a mások normatív rendszere felől érkező szankciókat.

Kollektív ágens esetén mondhatjuk, hogy az akkor autonóm, ha – valamilyen közösségi döntési eljárás során – saját maga által meghozott döntéssel saját magának ír elő kötelezettséget, és a kollektíva tagjai által elkövetett normasértéseket a kollektíva szankcionálja, illetve nem igazodik más, kívülről származó kötelezettséghez. Kollektív ágens jogi, illetve konvencionális normatív rendszert definiálhat magának. Előbbire az országok törvényhozási önrendelkezése lehet a példa (ezt nevezzük szuverenitásnak), utóbbira pedig valamely szervezet által birtokolt autonómia, mondjuk, az egyetemi autonómia.

A szervezetek, kollektív ágensek kapcsán mindig hangsúlyozni kell, hogy a szervezet létrehozásának indoka kijelöli a szervezeti működés számára releváns kötelezettségek halmazát, és csak ezekre vonatkozóan értelmes autonómiáról beszélni, mert ezen releváns kötelezettséghalmazon túl még sok olyan kötelezettséget lehet találni a szervezeti keretek között cselekvő emberek életében, amelyet valamely más, a szervezeten kívüli autoritás határoz meg. Ezekben a cselekvési tartományokban heteronómiáról kell beszélnünk. Egy törvényben megtilthatják a dohányzást az egyetemi épületekben, ami az egyetemi polgárok számára heteronóm igazodási kényszert teremt, de ettől még nem sérül az egyetemi autonómia elve, mert ezt a szabályozási elemet nem soroljuk az egyetemi élet releváns kötelezettségei közé tartozónak. Fontos megjegyezni azonban, hogy finanszírozási kérdések miatt egy szervezet sosem (vagy csak nagyon ritkán) lehet teljesen autonóm.

Most értünk el ahhoz a ponthoz, ahol megpróbálhatjuk megválaszolni a kérdést, hogy miként értelmezhetjük az autonómia fogalmát a gépi ágensekre vonatkozóan, az emberek kapcsán leírtakhoz képest. Amennyiben az autonómia alatt a teljes és kizárólagos önrendelkezést értjük, akkor ezt az elvárást nem terjeszthetjük ki a gépekre. Ennek két oka van: nem tudjuk pontosan definiálni az akarat fogalmát, és nincsen igazán erős tesztünk arra, hogy megállapíthassuk valamiről, rendelkezik-e akarattal. Az elsőt érzem fontosabbnak: az autonómia létezéséhez olyan minőségekre van szükség, amelyet az ember birtokol, de amelyekkel a gép most még biztosan nem rendelkezik, és egyelőre nem tudni, vajon fog-e valaha is rendelkezni velük.

Nem tudjuk megmondani, mi is az akarat, de emberként közös, szubjektíve folytonosan megélt tapasztalatunk az, hogy rendelkezünk vele. Mások cselekvései mögött is mindig ezt keressük, hogy megértsük, mit, mikor és milyen szándékkal cselekszenek. Az akarat minősége, a szabad akarat létezésének kérdése régóta vita tárgya. Egyesek szerint a szabad akarat csak illúzió (Wegner 2007), mások szerint nem. Vannak, akik a mérnöki tervezés számára is használható formulát ajánlanak az akarat fogalmának meghatározására. Michael Bratman szerint az akarat (vagy szándék) fogalmát úgy definiálhatjuk (Bratman 1987), hogy az akarat az, amikor az – egymással versengő – vágyak (desire) vagy célok (goal) közül „kiválasztunk" egyet, és elköteleződünk mellette (commitment). Az, hogy miként definiáljuk az akarat fogalmát, jelen gondolatmenet szempontjából másodlagos kérdés. A fontosabb kérdés inkább az, hogy miként tesztelhetjük, vajon létezik-e a vizsgált ágensnek (itt nyilván a gépnek) akarata. Mielőtt a válaszba kezdenék, meg kell még jegyezzem, hogy az autonómia értelmezéséhez nem csak az akarat fogalma szükséges, kell még valami más is: az önmagára irányuló akarat, az önmagának való parancsadás feltételezi az önreflexió képességét, és – vélhetőleg – az öntudat képességét. Az önreflexió, a tudat fogalma ugyancsak régóta vitatott kérdésköre a filozófiának, de ebben sem szükséges elmélyednünk. Amire itt választ kell keressünk, az az, hogy a gépek tudnak-e majd valamikor saját akaratból, belső késztetésre, nem mások külső parancsainak engedelmeskedve végrehajtani valamilyen aktust. Sokáig könnyű volt erre a kérdésre válaszolni, mert csak azt kellett mondani, hogy a számítógép az ember által megírt parancsokat hajtja végre. Ezen a tézisen akkor sem kellett érdemben változtatni, amikor lehetségessé vált, hogy a gép számára írt forráskódot egy másik gép írja, hiszen a gépet tekintve mindegy, honnan érkezik számára a külső utasítás. A gépi tanulás megjelenése jelentett fordulópontot ezen a téren, amikor kiderült, hogy a tanuló algoritmusok olyan megoldásokat alkalmaznak a saját repertoárjukból, és ennek megfelelően olyan eredményeket nyújtanak, amelyeket az ember előre nem programozott le, és utólag sem képes azokat jól megmagyarázni. Ennek alapján állíthatjuk-e, hogy ebben az esetben a gép már a saját akarata, a saját feje után ment?

Önmagában az a tény, hogy az ember nem képes „levezetni" vagy értelmezni az általa a gépnek adott utasításkészlet alapján a gépi által produkált eredményt, még nem jelenti a gépi akarat létezését. Az „if-then elágazásokban" való választás során valamilyen paraméter, vagy valamilyen elv, vagy a véletlen szerint dönthet a gép, hogyan megy tovább a kódban, és a gépi döntés az ilyen pontokban ember által előre nem látható eredményt ad, de ebben még nincs semmi, ami a gépi akaratra utalna, hasonlítana. Erre persze mondhatnánk azt, hogy épp ez az „előrejelezhetetlenség" a gépi akarat, a szabad döntés jele, mint ahogy a megengedő magatartás esetén embernél sem tudjuk előre, mi lesz a cselekvése iránya, végeredménye. Ez az előrejelezhetetlenség azonban nem annak tudható be, hogy a gép akaratát nem tudjuk kideríteni (mint egy ember esetében), hanem inkább saját tudásdeficitünknek. Ez egyébként igaz az emberi akaratra is, hiszen azt is visszavezethetjük rengeteg biológiai, fizikai állapotváltozás együttes hatásaként előálló eseménynek.

Akárhogy is definiáljuk az akarat fogalmát, és akárhogy is gondoljuk detektálhatónak az akarat létezését, ha az emberi értelemben vett akaratot tulajdonítjuk bármikor is a gépeknek, akkor az azzal jár együtt, hogy a gépek bármikor konfliktusba kerülhetnek az emberrel, hiszen a gépek szabad akaratukból kifolyólag mást akarhatnak éppen tenni, mint amit az ember elvárna tőlük. Ilyenkor mindig felmerül az akaratbefolyásolás kérdése. Ha a gépnek van szabad akarata, ami konfliktusba kerülhet egy ember akaratával, akkor mi lehet a konfliktus feloldásának módja? Természetesen az akaratbefolyásolás az emberek világában is létező jelenség, és az emberek közötti interakciókban is ugyanúgy kérdéses, hogy miként lehet biztosítani azt, hogy az egyik személy alávesse saját akaratát egy másik ember akaratának. Ennek többféle megoldása lehetséges a társadalomban. Ezek közül két társadalomtechnikát emelnék ki itt: meg lehet változtatni mások akaratát valamilyen kényszerítő eszköz alkalmazásával,

illetve engedelmességre lehet késztetni másokat kényszermentesen is. Az előbbit ragadhatjuk meg a hatalom, az utóbbit az uralom fogalmával (Weber 1987, Syi 2014).

Arra a kérdésre, hogy van-e tesztje az akaratnak, nem tudom a választ. Az akarat-teszt helyett azonban ki lehet próbálni egy másikat, egy akaratkonfliktus-tesztet, amely segíthet bizonyos kérdések eldöntésében. Ha a pórázra kötött kutyát húzom, hogy arra jöjjön, amerre én akarom, és ő ellenáll, akkor a megfeszülő póráz érvényes jelzése annak, hogy a kutyának van saját akarata. Ha meglát egy macskát és megkergetné, de én ezt nem akarom, akkor megint csak a póráz megfeszüléséből látszik, hogy ellentétes akarataink vannak. Ilyen esetekben két ágens között eltérő, tehát konfliktusban levő akaratot tapasztalhatunk, aminek vannak külső, tehát harmadik fél által is felismerhető jelei. A gépre vonatkozóan is feltehetjük a kérdést: mit tesz, ha létezik valamilyen előírás a cselekvésére nézve, vagy még pontosabban: megsértheti-e egy gép a számára előírt parancsot. A saját akarat létezésének a konfliktusok vizsgálatán keresztül – két erős próbája lehet:

- tiltás ellenére megcsinál valamit a gép, illetve
- kötelező parancs ellenére valamit nem hajt végre a gép.

Utóbbi esetben fontos szempont, hogy az utasítást nem valamely hiba miatt nem hajtja végre, hanem azért, mert mást akar csinálni. Az autonóm gépek megjelenésének jele az lesz, amikor egy gép valamilyen tiltó parancs ellenében mégis elvégez valamit, illetve egy kötelező modalitású utasítást nem hajt végre (a hibától eltekintve). Ma még nem ez a helyzet. Az természetesen előfordulhat, hogy bizonyos utasítások ellentmondanak egymásnak, és ilyenkor az egyik parancsot végrehajtva a másikat nem tartja be a gép, de ezt a „konfliktust” nem a gép maga oldja fel valamiféle mérlegeléssel, hanem a kódkészletében vannak úgy priorizálva az egyes utasítási szintek, hogy ezen jelzések mentén el lehet rendezni a konfliktusokat. Akkor beszélhetünk erős – kanti értelemben vett – autonómiáról, ha a gépek átmennek a fenti két akaratkonfliktusteszten. Mai tudásunk alapján ez a fajta gépi autonómia nem megvalósítható.

## *Eszközautonómia*

Az eddigiekben arról írtam, hogy a gépek autonómiáját nem tudjuk az emberi autonómia mintájára elképzelni, de fontos megjegyeznem, hogy az autonóm járművek, autonóm gépek fejlesztésének programja ettől még természetesen érvényes marad. Csak tudni kell, hogy még ha a legteljesebb autonómiáról beszélünk is a gépek esetén, másfajta értelemben kell ezt a fogalmat használni az emberi autonómiához képest.

Az autózás múltja, jelene és a szemünk előtt zajló fejlesztések nyilvánvalóan bizonyítják, hogy a járműveink egyre inkább önjáró eszközökké válnak. Az autó már a megnevezésével is utal az autonómia bizonyos fokára. A legelső autó is egyfajta automata volt: a motor vitte előre járműtestet, a vezetőnek csak a gázpedált kellett nyomnia, a kormányt kellett forgatnia. Az automata sebváltó, az elektromos ablakemelő, a sebességet tartó automat, a parkolóasszisztens mind olyan funkciók, amelyek egyre növelték a járművek automatizáltságát. Az önvezető járművek autonómiafokára már szabványokat hoztak létre. A SAE International nevű mérnökszervezet (Society of Automotive Engineers) 2014-ben rögzítette az elképzelését az önvezető járművek automatizáltságára, autonómiájára vonatkozóan (Adaptive 2015). Ez a hat szintű meghatározás az automatizáltság hiányától kezdve a részleges, feltételes, magas szintű automatizmusokon át a teljes automatizmust különíti el egymástól.

Az önvezető autók víziója arról szól, hogy az autóvezető megadja, hogy mi az úticél, milyen módon (milyen sebességgel és gyorsulással), milyen útvonalon kell haladni, mennyi idő alatt kell odaérni, mennyi pénzbe kerül az utazás, lehet-e fizetős autópályán menni vagy sem, milyen vezetési stílushoz (megfontolt, sportos, takarékos stb.) igazodjon a jármű stb. Ez egyfajta optimalizálási feladat, amit a gép jobban meg tud oldani, mint az ember. De ha azt

kérdezzük, hogy mi a járműhasználat célja, akkor arra nem nagyon felelhetünk mást, mint hogy eljutni valahonnan valahová[2], amihez meg kell adni egy úticélt, és ezt a célt mindig az ember határozza meg; ebben az értelemben az önvezető jármű sosem lesz autonóm. Amikor az autó elindult, akkor már lehet önvezető, hiszen minden döntést ő maga hoz meg, de az indulás előtti pillanatban, amikor a célt kell meghatározni, még az ember döntéstől függ minden. Az önvezető autó (vagy bármely más intelligens eszköz) nem lehet autonóm abban az értelemben, hogy a cselekvése releváns vagy végső értelmét adó célt maga határozhatná meg, de lehet autonóm abban az értelemben, hogy a releváns cél „elfogadása" után már minden más tevékenységet saját maga határoz meg.

A probléma pontosabb leírásához érdemes elkülöníteni a *célautonómia* és az *eszközautonómi*a (goal setting autonomy, illetve task autonomy) fogalmait egymástól. Akármilyen ágens cselekvéséről is beszélünk, igaz az a tétel, hogy a cselekvés mindig komponensekre bontható. Ha egy sötét szobában fényt szeretnénk, akkor ezt a célt többféleképp is elérhetjük, vagyis választhatunk az eszközök közül. Felkapcsolhatjuk a szobalámpát vagy egy állólámpát, meggyújthatunk egy gyertyát, használhatunk öngyújtót, bekapcsolhatjuk az okostelefonunk zseblámpáját stb. Bármelyik eszközt is választjuk, további teendőre van szükségünk. Ha a szobalámpát akarjuk felkapcsolni, akkor oda kell menni a kapcsolóhoz, és meg kell nyomni azt, de hangérzékelő kapcsoló esetén elég csak egyet tapsolni, és ég a lámpa. Ha az okostelefon zseblámpa funkcióját akarjuk használni, akkor elő kell kotorászni a telefont a táskánkból, meg kell keresni rajta a megfelelő ikont és kattintani kell egyet. Ezek a cselekvések a „fényt csinálni a szobában" cselekvés részeként, komponenseként értelmezhetők, és ilyenkor a cselekvés egésze felől tekintve még eszközként értelmezhető aktus már a részcselekvés céljaként minősíthető. A komponensek értelmezésekor a felsőbb szintű eszközök céllá változhatnak, ebben az értelemben a cél és eszköz szétválasztása mindig viszonylagos, hiszen a minősítés az elemzési szinttől függ. A cél és eszköz szétválasztása a cselekvés egészét tekintve lehet csak egyértelmű (utazás esetén: „hova menjek" a cél, az útirány, a sebesség, az úthossz, a fogyasztás stb. megválasztása már eszköz). Ebből a megközelítésből tekintve mondhatjuk, hogy az embereket célautonómoknak, míg a gépeket eszközautonómoknak tekinthetjük.

Ha az eddigiek alapján azt mondjuk, hogy nem tudjuk és/vagy nem akarjuk biztosítani a gépek emberi értelemben vett autonómiáját, és a gépeket az embert kiszolgáló szerepben szeretnénk tartani, akkor felmerül egy következő fontos kérdés: hogyan is kell értelmezni pontosan az ember-gép közti alárendelődési viszonyt. A következő fejezetben azt mutatom be, hogy az ember-gép viszony ugyanúgy paradoxonnal terhes, mint az ember-ember közti alárendelődési viszony. Ugyanez igaz, ha a gépek közti kapcsolatokat, utasítási-engedelmeskedési viszonyokat vizsgáljuk.

## *Kinek az ura, kinek a szolgája?*

Az ember és gép közötti viszony kapcsán gyakran idézik Isaac Asimovot, aki megfogalmazta a robotika három törvényét az intelligens gépek kötelmeire vonatkozóan (Asimov 2004).

1) *A robotnak nem szabad kárt okoznia emberi lényben, vagy tétlenül tűrnie, hogy emberi lény bármilyen kárt szenvedjen.*
2) *A robot engedelmeskedni tartozik az emberi lények utasításainak, kivéve, ha ezek az utasítások az első törvény előírásaiba ütköznének.*
3) *A robot tartozik saját védelméről gondoskodni, amennyiben ez nem ütközik az első vagy második törvény bármelyikének előírásaiba.*

[2] Elképzelhető persze, hogy az ember olykor magáért a vezetési vagy utazási élményért ül be az autóba, amikor az úticél érdektelen, de ezt az esetet itt most figyelmen kívül hagyhatjuk.

A három előírásból az első kettő vonatkozik az ember-gép kapcsolatra, a harmadik a gép saját cselekvésére irányul. A három törvény sorrendje fontos, a kisebb sorszámú parancsolat erősebb a többinél, ha tehát ütközés van két előírás között, akkor a kisebb sorszámú törvényt kell betartani. Ezen törvénykódex alapján a robot azt csinál, amit akar, amennyiben cselekvése nem kerül összeütközésbe a három parancsolattal. Mondhatnánk ezt részleges autonómiának, de teljesnek semmiképp. A harmadik törvény megtiltja a robot „önkárokozását", a második előírja az ember parancsainak végrehajtását, az első pedig kötelezi a robotot az embert érő kár megakadályozására. Az ember-gép viszonylatban a robot az „emberek utasításainak" van alárendelve. A robotika törvényei ellentmondanak a teljes és valódi gépi autonómiának. De a mesterséges intelligencia jövőjét boncolgatva nem is ez jelenti az igazi problémát. Ha elfogadjuk az embernek alárendelt gép vízióját, ha tudomásul vesszük, hogy a teljes autonómiát nem adjuk meg a gépek számára, akkor is maradnak még gondok.

Az első törvény második része megtiltja, hogy a robot passzív maradjon, amikor embereknek kárt okoznak. Ezt a károkozást nyilván más emberek tehetik csak meg, hiszen a robotok az első törvény első része alapján nem okozhatnak kárt embernek[3]. Ezen a ponton azonban gondot jelenthet a reciprocitás jelensége. A társadalom működésének, a kooperáció kialakításának és fenntartásának egyik alapmechanizmusa a reciprocitás, ami reaktív cselekvést jelent. A reciprocitás egyik formája a büntetés, a másik a jutalmazás. A büntetés egyidős lehet az emberi társadalom történetével. A büntetés valamilyen korábbi cselekvésre adott reakció, szándékos fájdalom- vagy károkozás a megbüntetett ember számára. A büntetés olyan társadalmi konstrukció, amit ember hoz létre azzal, hogy újfajta értelmezést ad a természeti szinten addig is létezett cselekvésének. Ugyanazt a fájdalom- vagy károkozást, ami – természeti szinten – addig is létezett és ami egyénileg rossz volt a büntetettnek, az emberi közösség átértelmezi, és azt mondja rá, hogy büntetésként konstruálva a közösségnek jó lesz. Ezáltal paradoxon keletkezik, az egyéni és közösségi értelmezések eltérnek egymástól, a két értékelés egyaránt választhatóvá válik. Az emberi társadalom egyik alapmechanizmusa tehát egy paradoxonnal terhes jelenség, és arra a kérdésre, hogy mit csináljon a robot, ha ilyen lát, melyik értékelköteleződéshez igazodjon, nincs egyértelmű válasz. Asimov „újszülött" robotjának az első sétája során ki kellene engednie a börtönökbe zárt embereket, hiszen azok ott kárt szenvednek, de ezzel nyilván a többi ember helytelenítését váltaná ki.

Az a tény, hogy Asimov idővel felvett egy negyedik parancsolatot is a „kódexébe" (Asimov 1993), amit – fontossága miatt – nulladiknak nevezett el, arra utal, hogy érzékelhette az első törvényben megbúvó problémát.

> *0) A robotnak nem szabad kárt okoznia az emberiségben, vagy tétlenül tűrnie, hogy az emberiség bármilyen kárt szenvedjen.*

Ha az emberiség asimovi kategóriáját az emberek közösségeként értelmezzük, akkor feloldódni látszik a fent bemutatott paradoxon, mert ez a kódex már megengedi azt, hogy az egyedi embernek okozott kárt a robot „eltűrje" a közösség érdekében, de persze ezzel kapcsolatban is felmerülhetnek problémák. Az emberiség ugyanis nem egyetlen közösség, a történelmünk fontos tapasztalata, hogy nagyobb közösségek egymással szemben álltak

[3] Ha feltételezzük is, hogy a robotok elméletileg nem okozhatnak kárt az embernek, ebből még nem következik az, hogy a gyakorlatban ez ne fordulhasson elő. Ahogy az ember, úgy a gép is hibázhat – a játékelméletből vett két gyönyörű hasonlattal élve – a reszkető kéz vagy a homályos látás miatt (Syi 2014). A homályos látás (gyengülő szem) metaforával azt a percepciós hibát írhatjuk le, amikor az ember vagy a gép nem jól ismeri fel, félreértelmezi a látottakat, és ezért másfajta motiváció alakul ki benne, és másként cselekszik, mint ahogy a helyes interpretáció esetén cselekedne, a reszkető kéz metaforával pedig a cselekvési hibára utalhatunk, amikor az ember vagy a gép nem pont azt a cselekvést hajtja végre, amit szeretett volna – akár saját hibájából, akár adott külső körülmények miatt.

különböző kérdések értelmezésekor. A kibővített kódex sem ad egyértelmű útmutatást arról, hogy kell-e bármit is tenni – mondjuk – a diktatúrákban börtönbe zárt ellenzéki aktivisták kiszabadítása érdekében, amire nyilván más választ ad a diktatúrában, illetve a „nem-diktatúrákban" élők közössége (most ne foglalkozzunk azzal a kérdéssel, vajon a diktatúrákban élő emberek egyetlen közösséget alkotnak-e). Az emberi társadalmakat megosztó értékelköteleződések mentén a robotoknak is ugyanúgy „választaniuk" kell, hogy melyik oldalra „állnak", viszont ezzel betarthatatlanná válik a robotkódex.

Ez a problémakör természetesen tovább mélyíthető és további kérdésekre lelhetünk, amelyek mindig az emberi társadalmak egyik kulcshelyzetéhez kapcsolódnak: ki, kinek, mikor, milyen feltételek mentén, milyen tartalmú utasítást fogalmazhat meg. Asimov novellájában egy ember utasítást ad egy robotnak.

> *„Megkeresed azt a robotot, és megparancsolod neki, hogy jöjjön vissza. Ha nem engedelmeskedne, erőszakkal hozod vissza."* (Asimov 2004)

A parancs olyan helyzetre utal, amelyben a robotnak egy másik robot cselekvését kellene befolyásolnia, és ez több kérdést is felvet. Először is: mikor, miért, miben engedelmeskedne az egyik robot a másiknak? Hogyan lehet szabályozni a robotok közti alá-, fölérendeltségi viszonyokat? Ezt mi, emberek, egymás közt úgy oldjuk meg, hogy engedelmességi viszonyokat teremtünk (és tartunk fent) azáltal, hogy cselekvési aszimmetriát érvényesítünk magunk között normák segítségével. Olyan uralmi normákat hozunk létre, amelyek adott helyzetre vonatkozóan parancsadási jogot adnak az egyik félnek, illetve engedelmességi kötelezettséget írnak elő a másik oldalon. Amikor a járványveszély idején a közintézmények kapujában mindenkit megállít egy ember, hogy megmérje a belépők testhőmérsékletét, akkor ebben a helyzetben az őr az, aki mindenki másnak előír valamit (nyújtsák oda a kezüket), amit mindenki engedelmesen végre is hajt. A következő pillanatban azonban az őrhöz oda léphet a felettese, és leválthatja őt, és onnantól fogva ugyanannak az embernek már senki nem köteles oda tartania a kezét. Ha pedig ugyanaz az őr az utcán sétálgatva akarná megmérni a járókelők testhőmérsékletét, visszautasításban részesülne, mert semmilyen – normatív – érvet nem tudna felhozni az igénye alátámasztására. Uralmi normából rengeteg van az emberi társadalomban. Ha ezt várjuk el a robotok közti cselekvésekkel kapcsolatban is, akkor létre kell hozni azt a robotjogrendszert, amely tételesen szabályozza azt, hogy milyen robotnak melyik másik robottal szemben, milyen feltételek mentén, milyen cselekvésekre vonatkozóan van utasítási joga. Ezzel persze megint szembesülhetnénk azzal a fentebb már említett dilemmával, hogy az országonként eltérő jogrendszerek más és más robotkódexet hoznának létre, és a robotoknak ugyanúgy igazodniuk kéne annak az országnak a jogrendszeréhez, ahol éppen tartózkodnak.

A fenti idézet kapcsán megválaszolandó másik kérdés az, hogy mit is jelent az, hogy a felszólított robotnak erőszakot kell alkalmaznia a neki nem engedelmeskedő robottal szemben. Ez az – emberi – utasítás megengedi az erőszak alkalmazását robotok között. Persze, az ember is alkalmaz erőszakot másokkal szemben (ezt a jelenséget nevezzük hatalomnak), de ezt a befolyásolási technikát a társadalomban elég régóta próbáljuk megakadályozni (normatív tiltásokkal). És mivel a robotika harmadik törvénye felszólítja a robotot saját védelmére, a roboterőszakra való emberi (asimovi) felszólítás a robotok közti harcot indukálja.

## Konklúzió

Tanulmányomban két tézis elfogadása mellett érveltem. Egyfelől azt állítottam, hogy az erős, emberi értelemben vett gépi autonómia megjelenésére és elterjedésére nem számolhatunk a belátható jövőben. Ez azonban egyáltalán nem is baj, hiszen az emberi értelemben vett autonóm gép nem lenne kívánatos. A tanulmány címében feltett kérdésre így az a válaszom, hogy a mesterséges intelligenciával rendelkező gép nem autonóm – ha az autonómia fogalmát

a kanti, erős értelemben definiáljuk. Számomra az embernek mellérendelt gép víziója egyelőre nem tűnik megvalósíthatónak. Mindebből természetesen nem következik az, hogy az embernek alárendelt gép vízióján alapuló, mesterséges intelligenciát használó gépek ne haladnák meg egyre több területen az ember képességeit. Az egyre intelligensebb alárendelt gépek azonban nem képesek az embert (vagy emberiséget) általában szolgálni. Az alárendelt gép víziójára támaszkodva nem lehet egyértelmű igazodási rendszert megfogalmazni a gépek számára, így a gépek utasítási, vezérlési jogosultságait be kell illeszteni a már létező társadalmi viszonyok közé. Ebből pedig az következik, hogy a gépek világában is fent kell tartani azokat a társadalmi különbségeket, amelyek eddig is voltak, és a jövőben is lesznek a különböző társadalmak között, illetve az egyes társadalmakon belül.

## IRODALOM


Adaptive (2015) *System Classification and Glossary*. https://www.adaptive-ip.eu/index.php/AdaptIVe-SP2-v12-DL-D2.1-System%20Classification-file=files-adaptive-content-downloads-Deliverables%20&%20papers-AdaptIVe-SP2-v12-DL-D2.1-System%20Classification.pdf (letöltve: 2021.03.21.)

Asimov, Isaac (1993) *Robotok és Birodalom*. Budapest, Móra Könyvkiadó.

Asimov, Isaac (2004) Körbe-körbe. In: Asimov, Isaac (2004) *Én, a robot*. Szeged, Szukits Könyvkiadó.

Boston Dymanics (2020*) Do You Love Me?* https://www.youtube.com/watch?v=fn3KWM1kuAw (letöltve: 2021.03.21)

Bratman, Michael (1987) *Intention, Plans, and Practical Reason*. Cambridge, MA, Harvard University Press.

Buss, Sarah (2002) Personal autonomy. *The Standford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/personal-autonomy (letöltve: 2021.03.21.)

Grady, Ken (2016) Checkers, Chess, Jeopardy, Go … Law. *Medium.* https://medium.com/the-algorithmic-society/checkers-chess-jeopardy-go-law-c42d20ae9910 (letöltve: 021.03. 21)

Hohfeld, Newcomb (2000) Alapvető jogi fogalmak a bírói érvelésben. In: Szabó Miklós – Varga Csaba (2000szerk.) *Jog és nyelv*. Budapest, Pázmány Péter Katolikus Egyetem.

Kant, Immanuel (1991) *Az erkölcsök metafizikájának alapvetése. A gyakorlati ész kritikája. Az erkölcsök metafizikája*. Budapest, Gondolat.

Lewis, David (1969) *Convention*. Cambridge, MA, Harvard University Press.

Millikan, Ruth G. (2008) Bioszemantika. In: Ambrus Gergely – Demeter Tamás – Forrai Gábor – Tőzsér János (2008szerk.) *Elmefilozófia*. L'Harmattan, 252–270.

Rousseau, Jean-Jacques (1964) *A magányos sétáló álmodozásai*. Budapest, Magyar Helikon.

Ryle, Gilbert (1974) *A szellem fogalma*. Budapest, Gondolat.

Searle, John R. (2000) *Elme, nyelv, társadalom. A való világ filozófiája.* Budapest, Vince Kiadó.

Syi (2014) *syi.hu/cse*. Budfapest, L'Harmattan – Könyvpont Könyvkiadó.

Weber, Max (1987) *Gazdaság és társadalom*. I. Budapest, Közgazdasági és Jogi Könyvkiadó.

Wegner, Daniel M. (2009) *A tudatos akarat illúziója*. Budapest, Kossuth Kiadó.

Wright, Georg Henrik von (1963) *Norm and Action*. London, Routledge and Kegan Paul.