A FOGOLYDILEMMA KITERJESZTÉSE

Tóth I. János

kandidátus/habilitált docens,
Szegedi Tudományegyetem Filozófia Tanszék
jtoth@philo.u-szeged.hu

A tanulmányomban arra a kérdésre keresek választ, hogy mi történik az egylépéses, kétszemélyes és kétstratégiás fogolydilemmában, ha eltekintünk az államhatalom külső szabályozásától, illetve ha realisztikus módon feltételezzük, hogy a játékosok nem feltétlenül becsületesek, illetve, hogy az erejük nem feltétlenül egyforma. Ebben az esetben a játékosoknak lehetőségük van a maximális nyereségre törekedni, amit csalással vagy erőszakkal tudnak realizálni. A modell kiterjesztése esetenként segít megoldani a fogolydilemma típusú interakciókat, másrészt a csalás és az erőszak olyan viselkedési formák, amelyek minden interakció esetében kialakulhatnak.

1. *A klasszikus fogolydilemma*

Az Albert Tucker megfogalmazta fogolydilemma (URL1), a 20. századi társadalomtudományok egyik legnagyobb hatású modelljévé vált (McCain, 2004, 9.). Jelentőségét jól illusztrálják Jon Elster (1976, 248.) szavai: *„a politikatudomány annak a vizsgálata, hogy milyen módon lehet meghaladni a fogolydilemmát."*

Az ismert történet szerint az ügyész vádalkut ajánl az egymástól izolált gyanúsítottaknak. A feleknek két döntési lehetőségük van: a vallomástétel (D), és a bűncselekmény tagadása (C). Ez négy különböző kimenetelhez vezethet: DC=T; CC=R; DD=P; CD=S. A fogolydilemma interakciójában a játékosok preferenciasorrendjét a következő egyenlőtlenségek jellemzik: T>R>P>S és 2R> T+S.

A modern játékelmélet terminológiája szerint a klasszikus fogolydilemma kétszemélyes, kétstratégiás egylépéses, nemkooperatív interakció. Ebben az interakcióban a játékosok Nash-egyensúlyi (URL2) (sőt egyben domi-

		2. gyanúsított	
	(döntés)	tagad (C)	vall (D)
1. gyanúsított	tagad (C)	R = -1**, R = -1**	S = -10, T = 0***
	vall (D)	T = 0***, S = -10	P = -6*, P = -6*

1. táblázat • Fogolydilemma. A cellákban az első szám az 1. gyanúsított, a második szám a 2. gyanúsított börtönbüntetését mutatja. Tekintve, hogy büntetésekről van szó, ezért a kimeneteleket negatív számok jelentik. A klasszikus fogolydilemmában racionálisan csak a következő szimmetrikus kimenetel várható: *= a játék egyensúlypontja; ** = kölcsönösen előnyös kimenetel. A kiterjesztett fogolydilemmában a maximális nyereség (***) megszerzése is lehetségesé válik.

náns[URL3]) stratégiája a vallomástétel (D), amit szoktak individuálisan racionális viselkedésnek is nevezni. A dilemma abból fakad, hogy az egyensúlypont egyben egy Pareto-inferior állapot, ahol a klasszikus (benthami) utilitarista jóléti függvény (URL4) értéke a minimális [$U_1(=P_1) + U_2(=P_2)$ =-12 év].

Anatol Rapoport (1965) behatóan elemezte a fogolydilemma-interakcióban a kooperálást és a dezertálást. Megállapította, hogy a kommunikáció lehetősége növeli a tagadás (C) arányát. A kooperáló viselkedést kollektíven racionális viselkedésnek nevezi, mivel ebben az esetben a legnagyobb a közös nyereség, pontosabban a legkisebb a közös veszteség [$U_1(=R_1) + U_2(=R_2)$ =-2 év]. Ugyanakkor az ún. kooperatív játékelméleti (URL5) kutatások rámutattak, hogy a tagad–tagad (CC) kimenetel instabil Pareto-optimális (URL6) kimenetelt jelent, amelynek a kooperatív játékelmélet terminológiája szerint üres a *magja* (core) (URL7). (A mag az olyan kimenetelek összességét jelenti, amelynek realizálásában minden játékos érdekelt.) Tehát pusztán az önérdekkövetés alapján a játékosok nem tudnak egyezséget kötni egymással.

Itt kell megjegyezni, hogy játékelméleti szempontból zavaró a tagadást kooperatív viselkedésnek nevezni, ugyanis a fogolydilemma eredendően nemkooperatív játék, ahol a játékosok között nincs egyeztetés.[1] Álláspontom szerint ezt a terminológiai nehézséget csak úgy tudjuk feloldani, ha különbséget teszünk a formális vagy procedurális és a szubsztantív kooperáció fogalma között. *Procedurális* kooperációról beszélhetünk, ha a feleknek a döntést megelőzően módjuk van arra, hogy megbeszéljék a döntési helyzetet és keressék a konszenzus lehetőségét. Egy játékos *szubsztantív* értelemben kooperatív, ha olyan döntést hoz, amelyik a másik érdekeit is figyelembe veszi. A klasszikus fogolydilemma esetében nincs mód a procedurális, csak a szubsztantív kooperációra.

Mancur Olson (1997) és Garret Hardin (2000) kutatásai alapján megállapítható, hogy a személyek számának növekedése radikálisan növelte a dezertáló viselkedés arányát a sokszemélyes (és egylépéses) fogolydilemmában. Ezen kutatások a következő fogalmakhoz vezettek: közjavak tragédiája *(tragedy of the commons)*, kollektív cselekvés *(collective acton)*, potyautas *(free rider problem)*, társadalmi dilemmák *(social dilemma)*.

Robert M. Axelrod (1984) a többlépéses fogolydilemma *(iterated prisoners' dilemma)* sajátosságait vizsgálva megállapította, hogy a döntési helyzet ismétlődése növeli a kooperálás lehetőségét. Kutatása kimutatta egy feltételesen együttműködő (*tit for tat* [URL9]) stratégia sikerét egy determinált környezetben.

2. *A fogolydilemma interakció kiterjesztése*

A fogolydilemma eredeti története szerint a gyanúsítottakat külön cellába helyezik, ami nyilvánvalóan keretek közé szorítja az interakciót. Felmerül a kérdés: mi történik, ha a modellt felszabadítjuk e korlátozás alól; vagyis a gyanúsítottak egy közös térbe kerülnek.

Első lépésben tekintsük el az őrök szabályozó hatásától, s tegyük fel, hogy a gyanúsítottakat egy táplálékban bőséges kis szigetre viszik, amely fölött egyetlen állam sem gyakorol szuverenitást. Filozófusok ezt a helyzetet *természeti állapotnak* (*state of nature* [URL10]) nevezik, ahol a szereplők bármit büntetlenül megtehetnek egymással szemben. A természeti állapot Thomas Hobbes szerint hadiállapot, de más filozófusok szerint (Locke, Spi-

[1] Erre dr. Pintér Miklós (Corvinus Egyetem Matematikai Tanszék) hívta fel a figyelmemet. (URL8)

noza) is olyan állapot, ahol könnyen kialakul az erőszak. William Golding *Legyek Ura* (1954), illetve Robert Merle *A sziget* (1962) c. regénye szintén a természeti állapotra utal, s mindkét történetben eszkalálódik az erőszak. Természeti állapotról vagy legalábbis ahhoz közeli állapotról beszélhetünk polgárháborúk, „törvényen kívül területek", amerikai vadnyugat, természeti katasztrófát követő társadalmi zavargások, szuverén nagyhatalmak relációjában stb. Azaz a természeti állapotban érvényesülő logika valamilyen mértékben mindig releváns az emberi viselkedés szempontjából.

Fentiekből következik, hogy az ügyész sem foglalkozik azzal, hogy mi történik a hatókörén kívül álló szigeten. Őt csak az eredeti bűncselekmény érdekli, és döntését megadott időpontokban (pl. minden este) leadható vallomások függvényében hozza meg a fent említett módon. Egy gyanúsított úgy tudja a tettét bevallani, hogy írásban lead egy ilyen jellegű vallomást, ezzel szemben a bűncselekményt tagadni kétféleképp is tudja: vagy lead egy ilyen értelmű nyilatkozatot, vagy egyáltalán nem ad le nyilatkozatot. Azaz: ha az ügyész csak egyetlen beismerő vallomást kap, akkor a vallomást tevőt szabadlábra helyezi.

Tehát az általunk vizsgált interakció mindenben megegyezik a klasszikus 2×2-es, egylépéses fogolydilemmával (*caeteris paribus* elv), kivéve a következő tényezőket.

a.) Az önérdekkövető játékosok szabadon érintkezhetnek egymással.

b.) A felek különböznek egymástól. A játékelmélet egyik alapvető előfeltevése, hogy az önérdekkövető és racionális játékosok minden szempontból azonosak, a valóságban azonban az emberek nem egyformák. Estünkben csak két különbséget vizsgálunk: a *becsületességet* és az *erőt*.

Mi történik a szigeten ilyen feltételek mellett?

3. Nyereségmaximum

Neumann János (1965, 121.) szerint egy interakcióban elvileg minden játékosnak az a célja, hogy az egész nyereség az övé legyen, de ezt a többi fél döntéseivel megakadályozhatja. Ezért ajánlja Neumann a minimax elvet. (URL11) Oskar Morgenstern (1959, viii,) szintén kizárta, hogy a játékosok az interakcióban elérhessék az abszolút nyereségpontot. S tény, hogy az abszolút nyereségmaximum a hagyományos játékelméleti keretek között nem szerezhető meg. A nemkooperatív játékelmélet keretei között ennek a kimenetelnek a megszerzése *fogalmilag* lehetetlen, hiszen a másik játékosnak ehhez egy nemegyensúlyi stratégiát kellene választani. A kooperatív játékelmélet keretei között pedig ennek a kimenetelnek a megszerzése *gyakorlatilag* lehetetlen. Ugyanis az első játékos hiába kéri meg a második játékost, hogy legyen szíves „tagadja a vádat" (C_2), hogy az első játékos maximalizálhassa a nyereségét, hiszen ekkor a második játékos nyeresége minimális lesz, és fordítva.

Álláspontom szerint a korábban megadott feltételek (a-b) mellett a maximális nyereség megszerezhető egy interakcióban. Létezhetnek ugyanis előzetes döntési folyamatok, amelyek eredményeképp egy játékos elveszti a döntési autonómiáját a másikkal szemben. A fölénybe kerülő játékos pedig arra is képes, hogy megszerezze az abszolút vagy a parciális nyereségmaximumot. Egy szigorúan önérdekkövető játékossal szemben csak kétfajta viselkedéssel lehet maximális nyereséget szerezni: csalással vagy erőszakkal. A csaló megtévesztéssel veszi rá a másikat arra, hogy önként tagadja a vádat (C). Erőszakos fél pedig erőszakkal veszi rá a másikat ugyanerre (C). További lehetőség nincs, hiszen valakit csak megtévesztéssel (nem erőszakkal) vagy

erőszakkal lehet rábírni arra, hogy olyat tegyen, ami nem biztos, hogy az érdekében áll.

A fogolydilemma esetében az abszolút nyereségmaximumot egy olyan aszimmetrikus kimenetel biztosítja, ahol az egyik (például az első) játékos beismeri a vádat (D), miközben a másik tagadja a vádat (C), ekkor az első játékos hazamehet, míg a második játékos 10 évet kap (D_1C_2=[T_1,S_2]=0,-10 év). Formálisan az abszolút nyereségmaximumot az ún. *maximax kritérium* (URL12) (sormaximumok maximuma) biztosítja. Az első játékos maximális nyeresége (T_1=0 év) azonban szükségszerűen együtt jár a második játékos minimális nyereségével (S_2=-10 év), vagyis a nyereségmaximalizáló törekvés az interakciót gyakorlatilag állandó összegű interakcióvá alakítja, ahol egy játékos vagy győz, vagy veszít.

A 2×2-es fogolydilemma esetében a csalás által elérhető parciális és az erőszak által elérhető abszolút nyereségmaximum egybeesik. Más interakciók esetében azonban ez a két kimenetel különbözhet egymástól. Általában a csalás csak parciális, míg az erőszakos viselkedés az abszolút nyereségmaximumot is képes biztosítani. A továbbiakban a nyereségmaximalizáló viselkedésnek ezt a két formáját vizsgálom (Tóth, 2010, 61.).

4. Csalás

Hogy megértsük a csalás (hamis kooperálás, szerződésszegés) logikáját, vizsgáljuk azt az esetet, amikor csak az egyik (például az első) játékos a csaló. A csaló úgy tesz, mintha kölcsönösen előnyös állapot kialakítására törekedne, és látszólag megegyezik egy ilyen kimenetelben (például CC=[R,R]=-1,-1). A „megegyezésből" adódóan a csaló előre tudja, hogy a másik hogyan fog viselkedni, és így – nem teljesítve az egyezségben vállalt kötelezettségeit – képes maximalizálni a saját nyereségét (Pl. $D_{csaló}C_{becsületes}$ = [$T_{csaló}$, $S_{becsületes}$]= $0_{csaló}$, $-10_{becsületes}$). Ugyanakkor a csalás csak akkor biztosítja a csalónak az abszolút nyereségmaximumot, ha a másik játékos becsületes, aki betartja a megállapodást. Ha mindkét játékos csal, akkor a játékosok csak az egyensúlyponthoz tartozó nyereséget szerzik meg.

A természeti állapotban nincs olyan külső erő, amely szankcionálná a csalást, vagyis a szerződésszegést. Szimultán játékban (URL13), ahol a játékosok egyidejűleg hozzák meg a döntéseiket, végső soron csak egyetlen módon lehet védekezni a csalás ellen, úgy, hogy nem kötnek megegyezést. Mellesleg ezért sem lehet a szerződésből levezetni a hatalmat, miközben a hatalomnak nyilvánvaló kötelessége a szerződések betartatása.

5. Erőszakos viselkedés

Az erőszakos viselkedés (URL14) meghatározó sajátossága fájdalom, sérülés és végső soron halál okozása. A fájdalom okozása esetenként öncélú, vagyis irracionális, gyakran azonban racionális eszköz a lehető legnagyobb nyereség (nyereségmaximum) elérése érdekében. Ezt tekintjük az erőszakos viselkedés végső (tartalmi) céljának. Formális szempontból az erőszak célja a másik akaratának az uralása (URL15). Az erőszakosan fellépő játékos engedelmességet vár a másiktól, különben valamilyen súlyos, az engedelmeskedéssel együtt járó hátránynál is lényegesen nagyobb hátránnyal (például erőszakos halállal) sújtja őt. Az erőszak formális logikáját a következő állítás fejezi ki: „megöllek, ha nem engedelmeskedsz az akaratomnak".

Világos, hogy az erőszakos viselkedés formális és tartalmi célja összefügg egymással, hiszen egy interakcióban csak az a fél tud maximális nyereségre szert tenni, aki rendelkezik a másik fél akarata fölött. Ezért az

erőszak elsődleges (vagy formális) célja a másik döntési autonómiájának felszámolása. Ennek megszerzése után a győztes uralja az általa legyőzött szolga döntési jogosítványait. Ebből pedig az következik, hogy a győztes mint úr képes nyereségmaximalizálni (a maximax döntési elvet alkalmazni) a vesztessel mint szolgával szemben.

Az erőszakos viselkedés sikere szempontjából meghatározó jelentősége van az erőnek. Egy játékos erejét egyrészt a természetes fizikai ereje, másrészt a nála levő fegyverek ereje együttesen határozza meg. Az erőről (f) feltesszük, hogy az értéke a teljes népességben véletlenszerűen oszlik el, például normális eloszlást követve. Ez azt jelenti, hogy a populációban viszonylag kevés ember van, aki az átlagnál sokkal erősebb, illetve sokkal gyengébb, és viszonylag sok olyan ember van, akinek az ereje az átlag közelében van. Tegyük fel, hogy a populáció tagjai közül bárki véletlenszerűen gyanúsítottá válhat. Ebből következően a két gyanúsított közötti erőviszony alapvetően a következő lehet: – az egyik fél sokkal (nyilvánvalóan) erősebb, mint a másik ($f_1>>f_2$); – az egyik fél kis mértékben erősebb, mint a másik ($f_1>f_2$); – a gyanúsítottak pont egyforma erősek ($f_1=f_2$).

Az erőszakos viselkedés kategóriáján belül különbséget kell tenni az erőszakkal való fenyegetés és az erőszak alkalmazása között. Az utóbbi elvileg lehet egyoldalú vagy kölcsönös, amit harcnak nevezek.

A *fenyegetés* (URL16) lényege, hogy a fenyegető a megfenyegetett személy számára súlyos hátrányt, kárt, veszteséget helyez kilátásba. Ez végső esetben az erőszakos halál (E_{death}). Lehetnek azonban olyan helyzetek (például munkahely, házasság esetében), ahol a kilátásba helyezett súlyos hátrány nem az erőszakos halál, hanem más jellegű kár (például a munkahely elvesztése vagy válás). Tehát a fenyegetés csak akkor hatásos, ha az így kilátásba helyezett veszteség sokkal nagyobb, mint az eredeti interakció legrosszabb kimenetele ($E_{death}<< S$). Ezt a súlyos hátrányt általában nem tartalmazza az interakciót jellemző kifizető függvény, ahogy például a fogolydilemma mátrixa sem tartalmaz olyan kimenetelt, mint a megsemmisülés. Ugyanakkor az emberek – különösen rendkívüli helyzetekben – számolnak ezzel a kimenetellel.

A fenyegető viselkedés kétféleképp is értelmezhető. Tekinthetjük úgy, mint az erőszakos viselkedésnek egy alacsony költségű formáját, ahol a fenyegetés sikertelensége esetében a fenyegető azonnal áttér a költségesebb harcra. Nevezzük ezt a viselkedési formát *harcos fenyegetésnek*. Másrészt a fenyegetést értelmezhetjük úgyis, mint egy tesztet, amely az erőviszonyok és az elszántságok felmérésére szolgál. Nevezzük ezt a viselkedési formát *tesztelő fenyegetésnek*. Ebben az esetben a fenyegetés során szerzett tapasztalattól függ, hogy a játékos harcolni fog-e vagy sem.

Egy fenyegetésre a megfenyegetett alapvetően kétféleképp válaszolhat: enged a fenyegetőnek, vagy szembeszáll a fenyegető játékossal. Ha enged a fenyegetésnek, akkor biztos elszenvedi a legkisebb nyereséget (S), ha szembeszáll a fenyegetéssel, akkor a harc kimenetele alapján nyer vagy veszít, de elvileg lehetséges a döntetlen is. Ebből következik, hogy más feltételek mellett következik be a támadó és a védekező erőszak, mivel a támadó inkább a lehetséges előnyökhöz, míg a védekező inkább a hátrányokhoz viszonyít. Az erőszak egyoldalú alkalmazása racionálisan viselkedő felek között nem alakulhat ki, mivel a megtámadott fél vagy megadja magát, vagy harcol. Ha egyiket sem teszi, akkor a támadó fél egyszerűen megöli őt.

Hatalmi harcról beszélünk, ha a támadó és a védekező is erőszakot alkalmaz. Ez a helyzet általában akkor alakul ki, ha a felek között nincs nyilvánvaló erőkülönbség. A hatalmi harc a társadalomtudományok egy fontos kategóriája. *„A gazdagságért, megbecsülésért, katonai parancsnokságért vagy más hatalomért folyó versengés viszályhoz, ellenségeskedéshez és háborúhoz vezet, mert minden versengő, hogy vágyát elérje, igyekszik versenytársát megölni, leigázni, félreállítani vagy visszaszorítani."* (Hobbes, 1999, 146.)

„A háború tehát erőszak alkalmazása, hogy ellenfelünket saját akaratunk teljesítésére kényszerítsük." (Clausewitz, 1961, 37.)

„Harcnak nevezünk egy társadalmi kapcsolatot, amennyiben a cselekvőt az a szándék vezeti, hogy a saját akaratát a másik vagy a többi féllel szemben keresztülvigye." (Weber, 1987, 64.)

A harc viszonylag gyorsan győztes-vesztes helyzethez vagy döntetlenhez vezet. Ha formálisan is modellezni akarjuk a harc kialakulását és lefolyását, akkor különböző modellek képzelhetők el.

- A legegyszerűbb modell szerint mindkét játékos pontosan ismeri a saját és a másik erejét, s a harc kimenetelét nem módosítja a véletlen. Ez esetben za erősebb fél győz, és a gyengébb szenved vereséget, ha pedig a felek pont egyforma erősek, akkor a harc végeredménye a döntetlen. Ebben a *determinista-objektivista* modellben a harc kimenetele csak és kizárólag a játékosok erejétől függ. Ebben a modellben az erősebb játékos, aki pontosan tudja, hogy ő az erősebb, nem köt kompromisszumot a gyengébbel, hanem a harcos fenyegetés után azonnal erőszakot alkalmaz. A gyengébb fél pedig akkor viselkedik racionálisan, ha támadás esetén azonnal feladja a harcot, feltéve, hogy a szolgaságot jobb állapotnak tekinti a megsemmisülésnél.
- A *determinista-szubjektivista* modell szerint minden játékos pontosan ismeri a saját erejét, de csak bizonyos valószínűséggel tudja megbecsülni a másik fél erejét. Ebben a modellben megnő a szerepe a *tesztelő fenyegetésnek*, hiszen így a játékosok pontosítani tudják a másik erejére vonatkozó becslésüket. A hozzávetőleg hasonlóan erős játékosok esetében ez a pontatlanság azt eredményezi, hogy mindkét félnek számolnia kell azzal a lehetőséggel is, hogy ő fog veszteni. Ezért a játékosok kis erőkülönbségek esetében lemondanak a kockázatos erőszakról és harcról, és inkább megpróbálnak megegyezni egy kisebb, de kölcsönösen előnyös kimenetelben.
- A *sztochasztikus-objektivista* modell szerint a játékosok pontosan ismerik az erőviszonyokat, ám a harc kimenetelét módosítja a véletlen. Azaz nagy valószínűséggel az erősebb fél győz, de kis valószínűséggel előfordulhat az is, hogy a gyengébb fél (Dávid esetenként legyőzi Góliátot). E modell szerint a harc mindig és mindenki számára magában hordja a vereség, illetve az erőszakos halál kockázatát.

Fontos kérdés, hogy egy játékos milyen negatív értéket tulajdonít a saját erőszakos halálának. Ha csak minimális értéket, akkor a játékos vakmerően (maximális bátorsággal) harcol minden olyan helyzetben, ahol erősebb. Az erős és vakmerő játékost jól példázza Akhilleusz, aki minden konfliktusban halálmegvető bátorsággal támad és harcol. S ez részéről egy racionális döntés, hiszen ő jóval nagyobb (várható) értéket tulajdonít az erőszakos halállal együtt járó örök dicsőségnek, mint a békés öregkorhoz társuló felejtésnek. A másik végletet az a (gyáva) játékos jelenti, aki végte-

len nagy veszteséget (negativitást) tulajdonít a saját erőszakos halálnak, ebből következően ha csak lehet, elkerüli a harcot. Arisztotelész a bátorságot a vakmerőség (irracionális bátorság) és a gyávaság (bátorság hiánya) között értelmezte. Az erőszakos haláltól, illetve a veréstől való félelem, amellyel mindkét félnek számolnia kell, önmagában növeli a megegyezés valószínűségét.

A *sztochasztikus-szubjektivista* modell szerint a játékosok csak bizonyos pontossággal (valószínűséggel) képesek felmérni az erőviszonyokat, és a harc kimenetelét a véletlen is befolyásolja. Hangsúlyozni kell, hogy itt két teljesen különböző valószínűségről van szó: az első egy szubjektív becslés a másik erejére vonatkoztatva, ami különböző pontosságú lehet, míg a második egy objektív bizonytalanság a harc kimenetelére vonatkozóan. A négy lehetséges modell közül itt a legvalószínűbb, hogy a felek nem erőszakosan, hanem kooperatív módon játsszák le az interakciót.

Ahogy nő a modellben a véletlen szerepe, úgy nő a bátorság jelentősége, vagyis annak, hogy egy játékos milyen mértékben hajlandó kockáztatni a legfőbb rossz, vagyis az erőszakos halál bekövetkezését. S épp a szubjektív és objektív bizonytalanságok miatt könnyen előfordulhat, hogy egy bátrabb, de objektíve gyengébb játékos megfélemlíthet, és így megadásra késztethet egy gyávább, de objektíve erősebb játékost is.

Minden viselkedési formának költsége van, amit szintén figyelembe kell venni. Hipotézisem szerint a nemkooperatív viselkedésnek (C_{NC}) a legkisebb a költsége, ezt követi a tisztességes és tisztességtelen tárgyalás (C_C), majd az erőszakkal való fenyegetés (C_T), és a legköltségesebb viselkedési forma a harc (C_F). A harc költsége (C_F) szűkebb értelemben csak a tényleges költségekre utal, míg szélesebb értelemben mindenféle kárt és kockázatot is magába foglal (mint például vereség, súlyos sérülés, erőszakos halál).

A fogolydilemma-interakció lefolyása tehát attól is függ, hogy hogyan viszonyulnak egymáshoz a különböző kimenetelek, illetve a költségek. Ha a kimenetelek jelentéktelenek (pl. a büntetés nem években, hanem percekben értendő), akkor nyilvánvalóan nem érdemes tárgyalni, csalni, illetve erőszakot alkalmazni. Tehát ha az interakció nyereségei (vagy veszteségei) nem jelentősek a többi lejátszási forma (kommunikáció, csalás, erőszak) költségeihez képest, akkor a játékosok számára az interakció leghatékonyabb lejátszását egyensúlyi stratégia követése jelenti (DD= [P, P] = -6, -6 perc). Lehetnek olyan nyereségek, ahol már érdemes vállalni a tárgyalás költségét, viszont még nem érdemes a költséges erőszakot alkalmazni. Végül lehetnek olyan interakciók, ahol a maximális nyereség (esetünkben a minimális vesztesség) megszerzését még harc árán is érdemes megkockáztatni.

6. A kiterjesztett fogolydilemma lefolyása

A fenti elemzést felhasználva röviden foglaljuk össze, hogy mi történik egy természeti állapotban levő fogolydilemma-interakcióban, ahol a játékosok becsületesség és erő szempontjából is különbözhetnek egymástól. Először is azt kell hangsúlyozni, hogy a különböző paraméterek (várható nyereségek, költségek, erő stb.) függvényében az interakció különbözőképp alakulhat. Bár itt és most hangsúlyozottan egy egylépéses interakcióról beszélünk, ennek mégis van egy időbeli lefolyása. Korlátozzuk a figyelmünket arra az esetre, amikor a várható nyereségek (illetve veszteségek) sokkal jelentősebbek, mint a harc költségei. Ebben az esetben az interakció várható lefolyását az *1. ábrán* látható séma mutatja.

Ahogy az ábra is mutatja, ebben az esetben a viselkedés egyre erőszakosabbá válik. Ugyanakkor ez a folyamat nem minden esetben jut el a harcig. Speciális esetben a feleknek sikerül már korábban lezárniuk a konfliktust.

Először a gyanúsítottak *tárgyalnak* egymással. Ez (vagyis a procedurális kooperáció) alacsony költségű konfliktusmegoldási módszer, tehát bármi is egy játékos végső célja, mindig érdemes megpróbálni azt tárgyalásos úton megszerezni. A tárgyalás célja lehet egy kölcsönösen előnyös kimenetelben (CC = [R,R] =-1, -1) való megegyezés. A tárgyalás elvileg végződhet az egyik fél maximális nyereségével (T=0 év) és a másik fél minimális nyereségével (S=-10 év). Ez a kimenetel akkor következik be, ha az egyik játékosnak, a csalónak sikerül becsapnia a másikat, a balekot.

Tehát a kiterjesztett fogolydilemma lehetőséget ad a procedurális kooperációra, amely elősegíti a szubsztantív kooperációt. Ugyanakkor ez a helyzet lehetőséget ad a nyereségmaximalizáló viselkedésre és ezzel párhuzamosan kiteszi a feleket a másik nyereségmaximalizáló viselkedésének is. Az adott feltételek mellett a „kölcsönösen előnyös kimenetel" csak akkor várható, ha egyik fél sem tekinti magát jelentősen erősebbnek, illetve ha a játékosok ki tudják zárni a csalás lehetőségét.

A *fenyegetés* átmeneti megoldás a tárgyalás és a harc között. A fenyegető játékos megmutatja erejét, és az erőszak alkalmazása melletti eltökéltségét. Ha a két fél között nyilvánvaló erőkülönbség van, akkor a gyengébb félnek érdemes harc nélkül behódolni; s ekkor az interakció kimenetele az egyoldalú dezertálás (DC=[T, S] = 0, -10 év). A fenyegetés azonban nem feltétlenül vezet erőszakhoz. Például egy hetvenkedő játékos definíciószerűen fenyegetéssel indít, ám az ellenállást tapasztalva visszatér a kooperációra (CC=[R, R] = -1, -1 év). (A hetvenkedő viselkedés fogalmát a többlépéses fogolydilemma modelljéből ismerjük, de ebben a folyamatban is értelmezhető.)

Ha a megfenyegetett nem engedelmeskedik, akkor a két fél közötti *hatalmi harc* alakul ki. A harc pedig vagy győztes-vesztes helyzethez vagy döntetlenhez vezet. Általában az erősebb fél győz, és a gyengébbnek fél szenved vereséget, aminek eredménye az, hogy a harc győztese vallomást (D) tesz, míg a harc vesztese kénytelen tagadni a vádat (C), s így a győztes azonnal kiszabadul, míg a vesztes tíz év börtönbüntetést kap. Ha az interakció döntetlennel végződik, akkor a feleknek nem marad más lehetőségük, mint hogy kölcsönösen vallomást tegyenek, és így mindketten hat-hat év börtönbüntetést kapnak.

A kiterjesztett fogolydilemmának van egy érdekes társadalomfilozófiai vetülete is, amennyiben az erőszak reális lehetőséget ad a fogolydilemma „meghaladására", majd a hatalmi struktúra korlátozása és demokratizálódása pedig lehetőséget ad az erőszak „meghaladására". A hatalmi harc nyilvánvalóan egy rossz társadalmi állapot, ugyanakkor ez egy olyan instabil helyzet, amely előbb-utóbb valamelyik oldal győzelmével stabilizálódik. A győztes-vesztes kimenetel nemcsak a győztes, de a közösség egésze számára is jobb kimenetelt jelent, mint a kölcsönös dezertálás. Ez már a kétszemélyes minitársadalom esetében is érvé-

TÁRGYALÁS$^{(\rightarrow R, R \text{ VAGY } T,S)}$ → FENYEGETÉS$^{(\rightarrow T,S \text{ VAGY } R, R)}$ → HARC → GYŐZTES-VESZTES (T,S)
VAGY DÖNTETLEN KIMENETEL → NEMKOOPERATÍV LEJÁTSZÁS (P, P)

1. ábra • A kiterjesztett fogolydilemma lejátszásának a sémája

nyes, hiszen a győztes-vesztes viszony kialakulása után klasszikus jóléti függvény értéke nagyobb [U_1 (=T) +U_2 (=S) = -10 év], mintha mindkét fél dezertálna [U_1 (=P) +U_2 (=P) = -12 év]. Minél nagyobb létszámú a közösség, annál egyértelműbb az úr-szolga viszony fölénye az anarchikus (természeti) állapottal szemben, mivel az úr minden szolgát kooperálásra kényszerít. Az úr-szolga viszonyban azonban az urat senki és semmi sem korlátozhatja, aki ezért továbbra is nyereségmaximalizáló, illetve erőszakos viselkedést folytathat. További előrelépést jelent az úr hatalmának megosztása és korlátozása, amely nemcsak békésebb, igazságosabb, de hatékonyabb társas állapotot is jelent [U_1 (=R) +U_2 (=R) = -2 év].

Kulcsszavak: *játékelmélet, maximális nyereség, csalás, erőszak, harc, természeti állapot, társadalomfilozófia*

IRODALOM

Axelrod, Robert M. (1984) *The Evolution of Cooperation.* Yale University Press, New Haven • http://www.sfs.uni-tuebingen.de/~roland/Literature/Axelrod(81)_the_evolution_of_cooperation.pdf

Clausewitz, Carl von (1961): *A háborúról* I. Zrínyi, Budapest

Elster, Jon (1976): Some Conceptual problems in Political Theory. In: Barry, Brian (ed.): *Power and Political Theory: Some European Perspectives.* Wiley, London, 245–270.

Gauthier, David (1969): *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes.* Clarendon Press, Oxford • http://books.google.hu/books?id=toEPO2Vy8sgC&printsec=frontcover#v=onepage&q&f=false

Hardin, Garrett (2000).: A közlegelők tragédiája. In: Lányi András (szerk.): *Természet és Szabadság. Humánökológiai olvasókönyv.* Osiris, Budapest, 219–231.

Hobbes, Thomas (1999): *Leviatán, vagy az egyházi és világi állam formája és hatalma.* (ford. Vámosi Pál – Ludassy Mária) Kossuth, Budapest

McCain, Roger A. (2004): *Game Theory. A Non-Technical Introduction to the Analysis of Strategy.* Thomson & South-Western

Morgenstern, Oskar (1959): Foreword. In: Shubik, Martin: *Strategy and Market Structure.* New York

Nash, John (1950): Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences of the USA.* 36, 1, 48–49. • http://www.pnas.org/content/36/1/48.full.pdf+html

Neumann János (1965): *Válogatott előadások és tanulmányok.* Közgazdasági és Jogi, Budapest

Olson, Mancur (1997): *A kollektív cselekvés logikája. Közjavak és csoportelmélet.* (ford. Csontos László) Osiris, Budapest

Rapoport, Anatol – Chammah, Albert M. (1965): *Prisoner's Dilemma,* The University of Michigan, Ann Arbor, MI • http://www.press.umich.edu/20265/prisoners_dilemma/?s=look_inside

Tóth I. János (2010): *Játékelméleti dilemmák társadalomfilozófiai alkalmazásokkal.* JATEPress, Szeged

Weber, Max (1987): *Gazdaság és társadalom.* (ford.: Józsa Péter) Közgazdasági és Jogi, Budapest

URL1: http://hu.wikipedia.org/wiki/Fogolydilemma

URL2: http://hu.wikipedia.org/wiki/Nash-egyensúly

URL3: http://www.gametheory.net/dictionary/DominantStrategy.html

URL4: http://en.wikipedia.org/wiki/Social_welfare_function

URL5: http://www.inf.unideb.hu/valseg/dolgozok/buraip/solymosi_jatekelmelet.pdf

URL6: http://hu.wikipedia.org/wiki/Pareto-hat%C3%A9konys%C3%A1g

URL7: http://en.wikipedia.org/wiki/Core_(game_theory)

URL8: http://sites.google.com/site/miklospinter/homeeng

URL9: http://en.wikipedia.org/wiki/Tit_for_tat

URL10: http://en.wikipedia.org/wiki/State_of_nature

URL11: http://hu.wikipedia.org/wiki/Minimax_elv

URL12: http://www.businessdictionary.com/definition/maximax-criterion.html

URL13: http://en.wikipedia.org/wiki/Simultaneous_game

URL14: http://hu.wikipedia.org/wiki/Erőszak

URL15: http://wikiszotar.hu/wiki/magyar_ertelmezo_szotar/Erőszak

URL16: http://www.kislexikon.hu/fenyegetes.html