Szociológiai Szemle 23(3): 30–50.

A reciprocitás rejtett mintázatai[1]

Syi
i@syi.hu

Beérkezés: 2013. 07. 08.
Átdolgozott változat beérkezése: 2013. 09. 28.
Elfogadás: 2013. 10. 07.

Összefoglaló: A kooperáció kialakulásában és fennmaradásában fontos szerepet tulajdoníthatunk a reciprocitáson alapuló társadalmi mechanizmusoknak. A direkt reciprocitás esetében négyféle cselekvési szabály, az indirekt reciprocitás esetében négyféle értékelési szabály létezik. A nyolc stabil, kooperációképes stratégia a cselekvés és az értékelés predikátumai, két ágensfogalom és a jóság kategóriája segítségével egyszerűen (ontológiailag laposan) formalizálható. A formális leírások alapján világossá tehető, hogy a cselekvési és értékelési szabályok párokba rendezhetők, melyeknek közös mintázatuk van. Ezt az összetartozást a büntetéshez való hasonló beállítódással magyarázhatjuk. Mivel a büntetés paradoxonából fakadóan a büntetés jelenségéhez szükségszerűen többértelműen viszonyulhatunk, így a tanulmányban elemzett négyfajta reciprocitáselv mindig egyszerre jelen van a különböző közösségekben, szubkultúrákban – noha változó társadalmi dinamikával.
Kulcsszavak: kooperáció, direkt reciprocitás, indirekt reciprocitás, cselekvési szabály, értékelési szabály, büntetés, másodrendű cselekvés, büntetés paradoxona

Az emberi társadalmak kooperáción alapulnak. Együttműködést persze találhatunk az állatvilágban is, de az ember mind minőségét, mind elterjedtségét tekintve sokkal magasabb szintű kooperációt valósít meg. Fel lehet tenni azt a kérdést, hogy miért is van ez, de még érdekesebb lehet annak megválaszolása, hogy milyen társadalmi mechanizmusok, cselekvési stratégiák képesek kialakítani és fenntartani a kooperációs mintákat olyan egyének között, akik egyébként folyamatosan erős kísértésben vannak a kölcsönös dezertálásra. A játékelmélet talán legismertebb játéktípusával, a fogolydilemmával ugyanis pont azt lehet bemutatni, hogy vannak társadalmi helyzetek, amikor – bár elvileg kívánatos lenne az együttműködés – a felek mégsem kooperálnak. A tényleges társadalmi gyakorlatban azonban olyankor is találkozhatunk kooperációval, amikor az az elmélet szerint nem lenne lehetséges. Az ellentmondás feloldható kétféle helyzet (játéktípus), az egyszeri és az ismételt játékok elkülönítésével. Sokszor kerülünk olyan – egyszeri – társadalmi helyzetekbe, amikor valóban működik a fogoly-

1 A cikk végleges változatának kialakításában nagyon sokat segített az egyik anonim bírálónak a szöveg formalizmusát érintő kritikája. Még többet köszönhetek Péli Gábornak, aki önzetlenül, nagy szakértelemmel segített a formulák szintaktikai hibáinak kijavításában. A cikk az OTKA K83887 sz. kutatása keretén belül született.

dilemma kölcsönös dezertálási logikája, de ha az ilyen helyzetek ismétlőd(het)nek, akkor már megjelen(het)nek a kooperációs minták is a felek között.

Az ilyen társadalmi helyzeteket – a játékelmélet eszközeire támaszkodva – az ismételt fogolydilemma-játékkal modellezhetjük (Axelrod 1984). Ezzel a játéktípussal a játékelméleten belül a *reciprocitás* jelenségét ragadhatjuk meg. A reciprocitás kategóriája a társadalomtudományok több ágában (szociológiában, antropológiában) már régóta kiemelt fogalomnak számít (Mauss 2004; Malinowski 1972; Sahlins 1973; Gouldner 1960), aztán a hetvenes évektől kezdve használni kezdték az evolúcióbiológia területén is (Trivers 1971; Alexander 1987). A nyolcvanas években fedezte fel magának a játékelmélet ezt a fogalmat, és azóta az evolúciós játékelmélet kutatói ontják magukból a különböző modelleket, szimulációkat (Sigmund 2010).

Trivers volt az, aki bevezette a *direkt reciprocitás* (DR) fogalmát, mint azt a mechanizmust, amely képes kialakítani és fenntartani a kooperációt két egymást ismerő ember (játékos) között. Ez utóbbi feltételt, hogy ti. ismerjék egymást az egymással társadalmi kapcsolatba került emberek, erős feltételnek kell minősítenünk, mert ez csak szűk körben, kis létszám esetén teljesül. A kooperáció létrejöttének magyarázatához a *cselekvési szabály* (*action rule* – ACR) fogalmát kell használnunk,[2] és arra a kérdésre, hogy milyen cselekvési szabályok lesznek sikeresek (kooperációképesek), a választ az *ismétlés* (*repetition*) mozzanatában találhatjuk meg (Sigmund 2010).

Alexander javasolta, hogy amikor már túl nagy a csoport, a társadalom mérete, és így a tagok már nem ismerik, nem ismerhetik egymást, az ilyen feltételek között kialakuló széles körű (*large-scale*) kooperáció jelenségére alkalmazzuk az *indirekt reciprocitás* (IR) fogalmát (Alexander 1987). Néhány korai kezdeményezés (Boyd – Richerson 1989) után a Sigmund–Nowak szerzőpáros volt az, aki megalkotta az indirekt reciprocitás – máig használt – formális játékelméleti modelljét (Sigmund – Nowak 1998; Nowak – Sigmund 1998). Az indirekt reciprocitás magyarázatához már nem elégséges a cselekvési szabály fogalma, szükség van felvenni mellé az *értékelési szabály* (*assessment rule* – ASR) kategóriáját is. Az értékelési szabályok segítségével lehet a társadalom, a csoport tagjainak *reputációját* megállapítani, változtatni. Mivel a tagok (játékosok) már nem ismerik egymást, nem tudhatják, ki hogyan viselkedett a korábbiakban, ezért, amikor dönteniük kell arról, hogy kooperáljanak-e a partnerükkel vagy sem, csak a reputációkra vonatkozó információ adhat fogódzót a döntéseikhez.

Mindkét reciprocitás modellezése esetén fontos feltétel, hogy a cselekvőknek csak kétféle cselekvési lehetőség adott: vagy kooperálnak, vagy dezertálnak (vagyis nem kooperálnak), amikor a kooperálás valami jót, a dezertálás valami rosszat jelent. Sahlins óta *pozitív reciprocitásról* beszélünk, ha valaki jóra jóval, vagyis kooperációval válaszol, míg *negatív reciprocitás* esetén valaki rosszra rosszal, vagyis dezertálással felel (Sahlins 1973).

2 A szabály és a stratégia terminusait itt egymással felcserélhető módon használom.

Direkt reciprocitás

A tanulmányban bemutatandó formulákhoz egy elsőrendű nyelvet használok (Ruzsa 2000), amire érvényes a predikátumkalkulus. Az L nyelv az alábbiakból áll:
L = {∧, →, ¬, ↔, =, ≠, VALUE, DO, *Good, agent, Ego, Alter,* ACR, ASR, G, B, C, D, TFT, CTFT, PAVLOV, GT, IS, SS, SJ, SH},

ahol:

∧	konjunkció logikai művelet (és)
→	kondicionális logikai művelet (ha, ... akkor)
¬	logikai tagadás művelet
↔	bikondicionális logikai művelet (akkor és csak akkor)
=	egyenlőség (kétargumentumú predikátum)
≠	egyenlőtlenség (kétargumentumú predikátum)
VALUE(x, t_i)	értékelés (kétargumentumú, kétértékű függvény): x értéke t_i időpontban
DO(*agent*, a, t_i)	ágens cselekvése (háromargumentumú predikátum): *agent* a-t cselekszi t_i időpontban
Good	az értékelési függvény pozitív értéke
agent	változó a cselekvőkre (ágensekre)
Ego	a „fokális" ágens (tulajdonnév)
Alter	a „másik" ágens (tulajdonnév)
ACR(n, *agent*, t_i)	cselekvési szabály (háromargumentumú predikátum): az n cselekvési szabály vonatkozik *agent*-re t_i időpontban
ASR(n, *agent*, t_i)	értékelési szabály (háromargumentumú predikátum): az n értékelési szabály vonatkozik *agent*-re t_i időpontban
G(x, t_i)	jó (kétargumentumú predikátum): x „jó" t_i időpontban
B(x, t_i)	rossz (kétargumentumú predikátum): x „rossz" t_i időpontban
C(*agent*, t_i)	*agent* kooperál t_i időpontban
D(*agent*, t_i)	*agent* dezertál t_i időpontban
TFT	Tit-for-tat cselekvési szabály (tulajdonnév)
CTFT	Bűnbánó tit-for-tat cselekvési szabály (tulajdonnév)
PAVLOV	Pavlov cselekvési szabály (tulajdonnév)
GT	Haragtartó cselekvési szabály (tulajdonnév)
IS	Image Scoring értékelési szabály (tulajdonnév)
SS	Simple Standing értékelési szabály (tulajdonnév)
SJ	Stern Judging értékelési szabály (tulajdonnév)
SH	Shunning értékelési szabály (tulajdonnév)

Ahhoz, hogy formalizálni tudjuk a legfontosabb cselekvési szabályokat, szükségünk lesz egy predikátumra, amivel a *cselekvés*[3] fogalmát ragadhatjuk meg:

$$\text{DO}(agent, a, t_i)$$

Itt elegendő számunkra, ha a cselekvés fogalmát három paraméterrel írjuk le. Egyfe-

3 A cselekvés fogalmát G. H. von Wright, S. Kanger, valamit N. Belnap és M. Perloff írásai alapján lehet jól leírni (von Wrigth 1963; Kanger – Kanger 1966; Belnap et al. 2001), de ezeket az elméleteket itt nem mutatom be.

lől meg kell adni a cselekvés ágensét. A jelen kontextusban a modellünkbe elegendő két ágenstípust felvenni (*Ego* és *Alter*).[4] Ezeket írhatjuk be a cselekvéspredikátum első változójába. Jeleznünk kell azt is, hogy a cselekvésnek mindig van valamilyen *a* tartalma, de arra egyáltalán nem lesz szükség, hogy kifejezzük a cselekvés tényleges tartalmát. Végül a harmadik paraméterrel utalhatunk arra, hogy melyik fordulónál tartunk a játéksorozatban.[5] Egy *cselekvési szabály* (ACR) mindig azt határozza meg, hogy egy ágensnek hogyan kell döntenie a játéksorozat valamely pontján.[6]

A kooperálás/dezertálás kettősség leírása végett be kell vezetni még egy kétértékű függvényt is, amivel az *értékelés* tevékenységét ragadhatjuk meg. A függvény két változójából az első paraméterrel azt a „dolgot" jelezhetjük, amire vonatkozóan az értékelést elvégezzük, a második változóval pedig az értékelés időpontját (a játéksorozat éppen adott fordulóját) fejezhetjük ki. Van tehát egy értékelésfüggvényünk:

$$\textsc{value}(x, t_i)$$

Az értékelési tevékenység modellezését megint csak érdemes leegyszerűsíteni. Mondhatjuk azt, hogy egy végletekig egyszerű, fekete-fehér világot feltételezünk, ahol az értékelések minősége csak kétféle lehet: vagy *jó* (*Good*) vagy *rossz* (*Bad*). Ha a *jó* fogalmát definiálatlan függvényértékként vesszük fel a leíró logikai nyelvünkbe, akkor a $\textsc{value}(x, t_i)$ értékelési függvény segítségével mind a *jó*, mind a *rossz* predikátumát – $G(x, t_i)$-t, illetve $B(x, t_i)$-t – definiálhatjuk a nyelven belül. Ezt úgy tehetjük meg, hogy kihasználjuk az értékelés dichotóm jellegét:

$$G(x, t_i) \leftrightarrow \textsc{value}(x, t_i) = \mathit{Good}$$
$$B(x, t_i) \leftrightarrow \neg(\textsc{value}(x, t_i) = \mathit{Good})$$

A természetes nyelvérzékünkhöz közelebb hozhatjuk a formális nyelvünket, ha az értékelő függvényünk számára felveszünk egy másik állandó függvényértéket is, hogy egyszerűbben is kifejezhessük a *rossz* értéket:

$$\neg(\textsc{value}(x, t_i) = \mathit{Good}) \leftrightarrow \textsc{value}(x, t_i) = \mathit{Bad}$$

Az eddigiekből már következik, hogy a fekete-fehér (dichotóm) világunkra igaz alábbi összefüggés:

$$B(x, t_i) \leftrightarrow \neg G(x, t_i)$$

Amennyiben az értékelési függvényünket alkalmazzuk a cselekvésekre, akkor a reciprocitáselméletek világán belül előforduló két lehetséges cselekvésminősítést

4 Nagyon szigorú tárgyalásmód esetén fel kellene még venni egy harmadik ágenstípust is, mégpedig egy kollektív ágenst, aki a közösségi értékelést végzi, de a kezelhetőség érdekében ettől itt eltekintek.

5 Az első fordulót t_1-gyel, az i fordulót t_i-vel, az i fordulót közvetlenül megelőző fordulót t_{i-1}-gyel jelöljük.

6 Az ismételt játékok összekötnek n darab játékot valamilyen játéktípusból, és a játékosoknak le kell játszaniuk az összes játékot egymás után. Az egyes játékokban elért eredményeket összegezve derül ki, hogy mely stratégiák gyűjtik össze a legtöbb pontot a játéksorozat végére. A sorozat egyes elemeire hivatkozhatunk a lépés, a forduló, a menet, a játék, a kör terminusokkal, a játékosok egyes játékokban meghozott döntéseire pedig utalhatunk a lépés vagy a cselekvés szavakkal is. A társadalmi emlékezet modellezésében addig megyünk el, hogy mindig csak a korábbi forduló adataira emlékeznek a játékosok, messzebbre nem látnak vissza a múltba. A korábbi, az előző terminusok tehát mindig csak az egy lépéssel korábbi játékra vonatkoznak.

már könnyen formalizálhatjuk. A dichotóm világban a cselekvésekre is áll az, hogy azok értékelése nem lehet más, mint *jó* vagy *rossz*.

$$\text{DO}(agent, a, t_i) \rightarrow (\text{VALUE}(a, t_i) = \text{Good} \leftrightarrow \neg(\text{VALUE}(a, t_i) = Bad))$$

Ha az *agent* cselekvő a t_i időpontban *a*-t cselekszi, akkor az a cselekvés értékelése t_i-kor 'jó', akkor és csak akkor, ha az *a* értékelése t_i-kor nem 'rossz'. Ugyanezt a már bevezetett $\text{G}(x, t_i)$ és $\text{B}(x, t_i)$ predikátumokkal is kifejezhetjük:

$$\text{DO}(agent, a, t_i) \rightarrow (\text{G}(a, t_i) \leftrightarrow \neg\text{B}(a, t_i))$$

Mondhatjuk ezt úgy is, hogy ebben a világban az *a* cselekvés vagy jó vagy rossz lehet, semmi más. A kétféle lehetséges cselekvésértékelés alapján már értelmezhetjük a kooperálás és dezertálás fogalmát is. A $\text{C}(agent, t_i)$ kooperálás a pozitívan, a $\text{D}(agent, t_i)$ dezertálás a negatívan értékelt cselekvést írja le:

$$\text{C}(agent, t_i) \leftrightarrow \text{DO}(agent, a, t_i) \wedge \text{VALUE}(\text{a}, t_i) = Good$$
$$\text{D}(agent, t_i) \leftrightarrow \text{DO}(agent, a, t_i) \wedge \text{VALUE}(a, t_i) = Bad$$

Az eddigiekből nyilvánvalóan következik az, hogy a két predikátum egymás tagadásának felel meg:

$$\text{D}(agent, t_i) \leftrightarrow \neg\text{C}(agent, t_i)$$

Innentől fogva már belekezdhetünk annak a kérdésnek a megválaszolásába, hogy ti. milyen kooperációképes cselekvési stratégiák vannak. Első lépésben érdemes rögzíteni, hogy vannak *feltételes* és *feltétel nélküli* cselekvési szabályok. A feltétel nélküli szabály két legnyilvánvalóbb példája a mindig kooperáló *AllC,* illetve a mindig dezertáló *AllD* stratégia. A játékelméleti szimulációk során az az – egyébként nem meglepő – eredmény adódott, hogy az *AllC* szabály sosem képes jó eredményeket elérni egy játéksorozatban, ezért a kooperációs mintát sem képes elterjeszteni a közösségen belül. Az *AllD* szabály lényegéből fakadóan kooperációellenes, amivel kapcsolatban inkább az a kérdés, hogy egyáltalán vannak-e olyan cselekvési szabályok, amelyek pont az *AllD* stratégiával szemben képesek a kooperációt kialakítani. Ez a kérdés azért megalapozott, mert a reciprocitáselméletek központjában az ismételt fogolydilemma-játékok állnak, és az ezzel kapcsolatos szimulációk nagy kérdése mindig az, hogy meg lehet-e haladni az egyszeri fogolydilemma-játékoknak azt a fontos tanulságát, miszerint az ilyen helyzetekben a kölcsönös dezertálás a domináns magatartás. A feltételes stratégiák esetén a feltételek – a jelen kontextusban – mindig valamely előző fordulóbeli eseményre, cselekvésre vonatkozhatnak. Ez két dolgot jelent. Egyfelől a feltétel fogalmát nagyon szűk értelemben használjuk, másfelől a feltételes stratégiák mindig másodrendű fogalmakkal ragadhatók meg. A cselekvési stratégiák tárgyalásához az elsőrendű logika mellett deontikus logikára is szükség van (von Wrigth 1963), hogy jelezni tudjuk a szabályokban tetten érhető kötelezés mozzanatát. Ezt a nyelvet nem emelem be a leírásba, elégnek tartom itt jelezni, hogy a cselekvési szabályok leírásához a deontikus

logika nyelvéből használnunk kell a kötelezőséget kifejező '**O**' deontikus operátort. Az alábbi formulával azt fejezhetjük ki, hogy az adott ágens számára az *a* cselekvést írja elő (a t_i időpontban) az adott cselekvési szabály (**O**: obligatory – kötelező).

$$\mathbf{O}\text{DO}(agent, a, t_i)$$

Arra a kérdésre keresve a választ, hogy melyek a kooperációképes feltételes stratégiák, a szimulációs számítások során négyféle cselekvési szabályt neveztek meg az idők során. Vegyük sorba mindegyiket![7]

(i) A leghíresebb kísérlet az a Robert Axelrod nevéhez fűződő számítógépes torna (Axelrod 1984), amelynek során egymással játszattak különböző cselekvési stratégiákat szimuláló programokat egy ismételt fogolydilemma-játéksorozatban. Ezen a tornán tűnt fel az Anatol Rapoport által benevezett stratégia, a *Tit-for-tat* (*tft*) szabály. A tornát megnyerő stratégia, amely tehát képes volt a kooperáció kialakítására, fenntartására és elterjesztésére, nagyon egyszerű, két részszabállyal leírható: ez első lépésben kooperálást ír elő, a többiben pedig az ellenfél előző fordulóbeli lépésének ismétlését. Ez a stratégia a direkt reciprocitás prototipikus szabálya abban az értelemben, hogy a többi stratégia is ezt a szabályt variálja valamilyen módon. A *Tit-for-tat* szabály ősrégi, bárhova nézünk a történelemben, minden kultúrában és szinte minden időben megtalálhatjuk ezt a cselekvési mintát. A „szemet szemért, fogat fogért" vagy az „aki bottal köszön, annak doronggal felelnek" mondások a negatív reciprocitásra, a „jótett helyébe jót várj" a pozitív reciprocitásra, míg a „kölcsön kenyér visszajár" vagy az „amilyen az adjonisten, olyan lesz a fogadjisten" mondások a reciprocitás mindkét ágára ráhúzhatók. Axelrod értékelése szerint ez a stratégia barátságos (mert jóindulatúan, kooperálással kezd), megtorló (mert a dezertálást azonnal megbünteti), megbocsátó (mert az első kooperálásra azonnal kooperálással felel, tehát „elfelejt" minden rosszat a múltból) és nem irigy (a tornák során egyetlen játékban sem kapott soha több pontot, mint az éppen adott ellenfele – mégis megnyerte a tornát). A stratégiát leírhatnánk egy összetett formulával (két kondicionális konjunkciójával), de érdemes kettéválasztani a két komponenst, mert az első összetevőt olyan félszabálynak minősíthetjük, amely az összes, később vizsgált cselekvési szabályra is alkalmazható. Ha felhasználjuk a korábban bevezetett predikátumokat, és a stratégiát *Ego* szemszögéből fogalmazzuk meg, akkor az első fordulóra a következő „állandó" szabályt vonatkoztathatjuk:

$$\text{ACR}(n, Ego, t_1) \leftrightarrow \mathbf{O}\text{C}(Ego, t_1)$$

A formula azt írja le, hogy az *n* nevű ACR cselekvési szabály a t_1 első fordulóban az *Ego* számára kötelezően előírja a kooperálást: $\mathbf{O}\text{C}(Ego, t_1)$. Ezt kivételszabályként is felfoghatjuk, hiszen erre „csak azért van szükség", mert az első fordulóban a főszabályt még

7 A négy szabályról bővebben írtam korábban (Syi 2008, 2012).

nem lehet alkalmazni. Azért nem, mert az első fordulóban nincs „előző forduló". Ezt viszont minden cselekvési szabály esetén érvényesítenünk kell, tehát bármit is írnánk be az *n* változó helyébe, mindig ugyanezt a kivételszabályt kéne alkalmaznunk az első fordulókra. Ezért a továbbiakban ezzel nem is kell foglalkoznunk, hiszen ez nem ad hozzá semmi különöset a cselekvési szabályok jellemzéséhez. A különbségek az első forduló után alkalmazandó szabálykomponensekben keresendők. Ezt a TFT nevű cselekvési szabály esetében *Ego* számára a következő módon írhatjuk le:

$$\text{ACR}(\text{TFT}, \mathit{Ego}, t_i) \leftrightarrow ((\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i)) \wedge (\text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i)))$$

A szabály szerint, ha az előző t_{i-1} menetben az *Alter* dezertált (D), akkor a vizsgált t_i fordulóban *Ego* számára a dezertálás van előírva ($\mathbf{O}$D(*Ego*, t_i)), míg *Alter* kooperálása (C) esetén *Ego*-nak is kooperálnia kell ($\mathbf{O}$C(*Ego*, t_i)). Ezt a leírást érdemes redundáns módon kibővíteni azért, hogy a későbbiekben jobban összevethetővé váljék a még bemutatandó további szabályok formuláival.

$$\text{ACR}(\text{TFT}, \mathit{Ego}, t_i) \leftrightarrow (\text{C}(\mathit{Ego}, t_{i-1}) \rightarrow (\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i) \wedge \text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i))) \wedge (\text{D}(\mathit{Ego}, t_{i-1}) \rightarrow (\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i) \wedge \text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i)))$$

Ezzel a hosszabb formulával ugyanazt a cselekvési tartalmat írjuk elő, mint az előbb, csak felvettünk ide egy redundáns feltételt. A kibővített képlet azért redundáns az előző alakjához képest, mert ugyanazt a cselekvést írja elő arra a két esetre, amikor az *Ego* az előző fordulóban kooperált, illetve dezertált, ami miatt fölösleges lenne ezt a feltételt figyelembe venni, ha az összehasonlíthatósági szempontokra nem figyelnénk.

(ii) R. Sugden javasolta a *Tit-for-tat* módosított változatát (Sugden 1986), amire ő T_1 szabályként hivatkozott. A későbbiekben az indirekt reciprocitás jelenségét elemezve *Standing* stratégiának nevezték el Sugden szabályát, majd Boyd ráakasztotta a *bűnbánó TFT* címkét (Boyd 1989; Panchanathan – Boyd 2003), illetve megint mások a *szigorú, de igazságos* (*Firm-But-Fair – FBF*) nevet húzták rá (Frean 1994; Hauert – Schuster 1998). Az utóbbi bizonyos szempontból nagyon találó, ráadásul széles körben használt kifejezés, ám én mégis a *bűnbánó TFT* (CTFT) megnevezést fogom alkalmazni rá, mert a bűnbánás mozzanata ragadja meg leginkább a stratégia legfontosabb vonását (Boerlijst et al. 1997). A szabály ugyanis azt írja elő, hogy ha az előző fordulóban az *Ego* kooperált (tehát jó volt), akkor a partner előző lépését kell tükrözni, ám ha *Ego* dezertált az előző menetben, akkor utána mindenképpen kooperálnia kell. A rosszat meg kell bánni, a rosszat ki kell javítani – ezt mondja a szabály. A hozzá rendelhető formula pedig így néz ki:

$$\text{ACR}(\text{CTFT}, \mathit{Ego}, t_i) \leftrightarrow (\text{C}(\mathit{Ego}, t_{i-1}) \rightarrow (\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i) \wedge \text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i))) \wedge (\text{D}(\mathit{Ego}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i))$$

Csak az egyes lépésekre figyelve ez a legkooperatívabb stratégia, hiszen három alkalommal írja elő a kooperálást (a lehetséges négyből).[8] A bűnbánásról szóló narratívák

8 Ez a szabályok táblázatos formájú reprezentációjából derül ki igazán, amit a tanulmány későbbi részében mutatok be.

hátterében vélhetőleg mindig ez a stratégia áll. Jézus híres mondásából, az „aki közületek nem bűnös, az vesse rá először a követ" (Jn. 8,7) felszólításból is ez a stratégia bontható ki: aki bűnös (mert korábban dezertált, tehát vétkezett), az most ne ítélkezzen, legyen bűnbánó, és kooperáljon.

(iii) Még a reciprocitáselméletek felfutása előtt, tehát azoktól még részben függetlenül vezették be a *Pavlov* (PAVLOV) -stratégia fogalmát (Kraines – Kraines 1989, 1993), amit később Rapoport *Együgyű* (*Simpleton*) szabálynak nevezett (Ridley 1997), aztán voltak, akik a *perfekt TFT* címkét használták rá (Imhofa et al. 2007), végül szélesebb körben elterjedt a *győztes csapaton ne változtass* (*Win-Stay-Lose-Shift* – *WSLS*) szabálynév erre a stratégiára (Nowak – Sigmund 1993; Imhofa et al. 2007). Ez a szabály egyfajta tanulási stratégiaként is értelmezhető, és a szimulációs számítások nagy része szerint ez rendelkezik a legjobb kooperációképességgel (Nowak – Sigmund 1993). A szabály formulája a következő:

$$\text{ACR}(\text{PAVLOV}, \mathit{Ego}, t_i) \leftrightarrow (\text{C}(\mathit{Ego}, t_{i-1}) \rightarrow (\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i) \wedge \text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i))) \wedge (\text{D}(\mathit{Ego}, t_{i-1}) \rightarrow (\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i) \wedge \text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i)))$$

A formula mutatja, hogy ez a legkomplexebb stratégia, mert a konjunkció mindkét tagjában a partner előző lépéséhez igazítja a saját, soron következő cselekvését. Arra az esetre, amikor *Ego* kooperált az előző fordulóban, ugyanazt írja elő, mint az előző két szabály (a partner lépésének tükrözését), ám ha az *Ego* rossz volt korábban, vagyis dezertált az előző lépésben, akkor a második részszabály a partner előző lépésének az ellentétét követeli meg. Ez a szabály „csavar" egyet a reciprok viszonyuláson, ennyiben eltér az egyszerű *TFT* stratégiától. A *Pavlov* szabály vélhetőleg ennek a mozzanatnak köszönheti a sikerességét. Míg a *TFT* csak a partnere korábbi lépését veszi figyelembe, addig a *Pavlov* szabály már figyelembe veszi saját cselekvését is.

(iv) A negyedi cselekvési szabály a *Grim Trigger* (GT) -stratégia (Friedman 1971). Ez a szabály nem megbocsátó, sőt inkább bosszúálló, *haragtartó* stratégia (néha hivatkoznak rá Friedman- vagy állandóan megtorló szabályként is). Miért? A szabály addig kooperál, amíg a partnere is ezt teszi, azonban az ellenfél első dezertálására azonnal és visszavonhatatlanul vég nélküli dezertálásra vált át. Elég egy hiba, egy dezertálás, és nincs visszaút, nincs megbocsátás. A formulája:

$$\text{ACR}(\text{GT}, \mathit{Ego}, t_i) \leftrightarrow (\text{C}(\mathit{Ego}, t_{i-1}) \rightarrow (\text{D}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i) \wedge \text{C}(\mathit{Alter}, t_{i-1}) \rightarrow \mathbf{O}\text{C}(\mathit{Ego}, t_i))) \wedge (\text{D}(\mathit{Ego}, t_{i-1}) \rightarrow \mathbf{O}\text{D}(\mathit{Ego}, t_i))$$

Nincs olyan ismert és használt közmondás, amely ezt a szabályt foglalná magába, de a 'haragtartó személyiség' kifejezés azt a beállítódást írja le, amely e stratégia mögött áll. Bár a tényleges társadalmi gyakorlatokban a 'nincs kegyelem' elvét fellelhetjük, ezt az elvet nem vagy csak nagyon ritkán emelték fel a követendő szabály rangjára. A játékelméleti szerzők között mégis voltak, akik a haragtartó stratégiát erősen kooperációképesnek gondolták azon az alapon, hogy nagy benne a fenyegetési potenciál: „Az ismételt játékokra vonatkozó úgynevezett hétköznapi teóriák

szerint, ha a következő fordulók valószínűsége elég magas, a kooperáció fenntartható az úgynevezett Trigger-stratégiákkal, amelyek azonnal vég nélküli dezertálásba csapnak át, ha a partnerük egyszer dezertál" (Nowak – Sigmund 2005). A későbbi szimulációk azonban ezt az elvárást nem támasztották alá.

(v) Az itt bemutatott négy cselekvési szabály került az elemzések fókuszába a direkt reciprocitás jelenségének modellezésekor. Ezek a stratégiák – bizonyos feltételek mentén, kisebb-nagyobb mértékben – képesek voltak a kooperációs minták kialakítására és fenntartására. Ha egy közös táblázatba rendezzük a négy stratégia jellemzőit, akkor érdekes hasonlóságokat és eltéréseket vehetünk észre. A formulákba foglalt szabályokat a táblázatos formában másként kell kifejeznünk. A két ágensnek minden egyes fordulóban két-két cselekvési lehetősége van (*C* vagy *D*), ebből következően négyféle világállapothoz lehet hozzárendelni a cselekvési stratégiák részszabályait: létezik a kölcsönös kooperáció (*CC*) és a kölcsönös dezertálás (*DD*), illetve van kétféle kevert eset (*CD*, illetve *DC*). Az egyes szabályokat jellemző feltételes cselekvésértékeket úgy írhatjuk be a döntési mátrixba, hogy jelezzük, melyik világállapothoz milyen cselekvést (*C*-t vagy *D*-t) ír elő a szabály *(1. táblázat)*.

1. táblázat: *A négy cselekvési szabály mintázata*

ACR ↓	$\text{DO}(Ego, a, t_{i-1}) \rightarrow$	C		D	
	$\text{DO}(Alter, a, t_{i-1}) \rightarrow$	C	D	C	D
Simple Tit-for-tat	$\mathbf{O}\text{DO}(Ego, a, t_i) \rightrightarrows$	C	D	C	D
Contrite Tit-for-tat		C	D	C	C
Pavlov		C	D	D	C
Grim Tigger		C	D	D	D

A négy stratégia közös vonása, hogy ha az *Ego* kooperált az előző fordulóban, akkor mindegyik szabály a partner korábbi lépését ismételteti meg. A különbség abban van, hogy mit írnak elő arra az esetre, amikor *Ego* előzőleg dezertált. Elméletileg négyféle ilyen eset létezhet, és az itt vizsgált négy cselekvési szabály pontosan ennek az elméletileg adott négyféle lehetőségnek feleltethető meg. Ez azt is jelenti, hogy a direkt reciprocitás koooperációképes cselekvési szabályait keresve nem találhatunk mást, mint a most bemutatott négy stratégiát.[9]

Indirekt reciprocitás

Az *indirekt reciprocitás* (IR) a széles körű kooperáció mechanizmusa, amikor a társadalom, a csoport mérete már akkorára nő, hogy a tagok (játékosok) nem ismerhetik valamennyien egymást. Sigmund és Nowak dolgozták ki az első formális modellt erre

9 Természetesen csak a modellezésnek ezen a szintjén. Ha bővítjük a modellünket, akkor – elméletileg – tér nyílhat új stratégiák számára.

a jelenségre (Sigmund – Nowak 1998; Nowak – Sigmund 1998). Az új modellben szükségessé válik a stratégia fogalmának bővítése. Míg DR környezetben a stratégia fogalma egyet jelentett a cselekvési szabály (ACR) fogalmával, addig az IR kontextusban a cselekvési szabály mellé fel kell venni egy új kategóriát, az *értékelési szabály (assessment rule* – ASR) fogalmát.[10] Egy ASR határozza meg a játékosok *reputációját,* amely azért szükséges, mert ha a játékosok nem ismerik egymást, tehát nem tudják, ki hogyan viselkedett korábban, attól még valahogyan dönteniük kell arról, hogy kooperáljanak-e vagy sem az éppen adott fordulóban.[11] A reputáció fogalma olyan értékfogalom, amelynek az általános VALUE(x, t_i) függvénynél jóval szűkebb az értelmezési tartománya, hiszen csak ágensekre vonatkoztatható. Ebben az esetben is igaz viszont, hogy a reputációnak kétféle értéke lehet (jó vagy rossz ágens), amit úgy fejezhetünk ki, hogy a G(x, t_i) „jó" és a B(x, t_i) „rossz" predikátumok esetében az x változóba az *agent* valamelyik értékét, tehát vagy az *Ego*-t vagy az *Alter*-t helyettesítjük.

Az értékelési szabály fogalmát megint csak célszerű valamilyen modális logikával megragadni. Az egyszerűség kedvéért azt mondom, hogy az értékelési tevékenységet értelmezzük kollektív cselekvésként, amit az adott fordulóban a *Donor* cselekvése alapján (és után) végez el a közösség, és az értékelési tevékenység eredményét (az új reputációt) rendeljük hozzá a Donorhoz, és jelöljük mindezt az '**A**' operátor bevezetésével az alábbi módon:

$$\mathbf{A}\{\text{VALUE}(agent, t_i) = Good\}$$

Az új **A** szimbólumnak a deontikus logika nyelvén azt az értelmezést adhatjuk, hogy a közösség számára az adott ágensre vonatkozó adott értékelést ír elő (a t_i időpontban) az adott értékelési szabály.[12] Meg kell még jegyezzem, hogy az értékelési szabály argumentumaként megadott ágens passzív szerepben van (rá vonatkozik az értékelés), míg a cselekvési szabály esetében az ágens aktív szerepet visz (a szabály az ő cselekvését írja elő). Bár fentebb azt jeleztem, hogy az indirekt reciprocitás leírásához két összetevőre kell bontani a stratégia fogalmát, a továbbiakban mégsem kell mindkét összetevővel foglalkoznunk. Az indirekt reciprocitás jellemzése során – a jelen elemzési szempontok szerint – elhagyhatjuk a cselekvési szabályokat, mondanivalónk kifejtéséhez elegendő csak az értékelési szabályokkal foglalkoznunk.[13] A kulcskérdés itt az, hogy miként lehet a reputációt meghatározni, mikor és hogyan lehet változtatni az ágensek minősítésén. Az értékelési szabályok úgy működnek, hogy – az ágensek cselekvését az egyes fordulókban figyelve – meghatározzák, hogy az *Ego* (IR-kontextusban a *Donor*)

10 Az értékelési szabály fogalmára használhatjuk még az értékelési stratégia, a mechanizmus és a módszertan kifejezéseket is.

11 Az indirekt reciprocitás körülményei közt a kooperálás segítségnyújtásnak, donációnak felel meg, míg dezertálásnak a segítségnyújtás megtagadása számít.

12 Ahogy már jeleztem, egy egzaktabb tárgyalás során szükségünk lenne valamilyen kollektív ágens fogalomra, de ezt most itt elhagyhatjuk.

13 Az elméletileg lehetséges sokféle cselekvési szabály közül a most bemutatandó kooperációkész stratégiákban mindig kétféle cselekvési szabály fordulhat csak elő, az ún. 'CO' (diszkriminatív kooperáló) és az 'OR' stratégia. Az indirekt reciprocitás jelenségének megértéséhez keveset ad a cselekvési szabály fogalma. Ezeket korábban máshol alaposabban bemutattam (Syi 2008).

reputációja hogyan változzon az alapján, hogy az *Ego* (*Donor*) kooperált vagy dezertált az *Alter*-rel (IR-kontextusban: a *Recipiens*-sel) szemben.

(i) Az első IR-modellben Sigmund és Nowak bemutatták az Image Scoring (IS) értékelési szabályt, amit magyarul „kép-mutató" (kép-pontozó) szabálynak mondhatnánk (Sigmund – Nowak 1998). A stratégia nagyon egyszerű (a *Tit-for-tat* cselekvési szabályhoz hasonlóan ez is elsőrendű szabály). A stratégia előírja, hogy induláskor a játékban (tehát az első fordulóban) mindenki jó (*G*) minősítést kap:

$$\text{VALUE}(agent, t_1) = Good$$

Ezután minden fordulóban értékelni kell a játékosok cselekvését. *Ego* reputációja jó (*G*) lesz, ha kooperált, rossz (*B*), ha dezertált. Ha formalizálni akarjuk ezt a szabályt, akkor az első lépésben ugyanúgy járhatunk el, mint ahogy azt a cselekvési szabályok leírásakor tettük. Elhagyhatjuk az első fordulóra vonatkozó fenti részszabályt (mindenki jó), hiszen ez mindegyik értékelési mechanizmusra érvényes lesz, valamint elhagyhatjuk a további fordulókat jelző feltételt is a formulákból. A kérdés az, hogy hogyan tudjuk megragadni azt, hogy az *Image Scoring* módszertan az *i.* fordulóban *Ego* cselekvése ($\text{DO}(Ego, a, t_i)$) alapján új reputációt ($\text{VALUE}(Ego, t_i)$) rendel *Ego*-hoz. Nos, azt a tényt, hogy ez a szabály a kooperáláshoz (*C*) jó (*G*), a dezertáláshoz (*D*) rossz (*B*) minősítést rendel, azzal fejezhetjük ki, hogy kijelentjük, a cselekvés és az ágens értékminősítése megegyezik:

$$\text{DO}(Ego, a, t_i) \rightarrow \text{VALUE}(Ego, t_i) = \text{VALUE}(a, t_i)$$

Ezt az összefüggést beemelhetjük az értékelési szabály leírásába úgy, hogy erre vonatkoztatjuk az értékelés szabály jellegét kifejező **A** értékelési operátort. Az így kapott formula ugyanúgy elsőrendű, mint a *Tit-for-tat* szabály formulája volt, és a további értékelési szabályokkal való összehasonlíthatóság kedvéért ezt is érdemes redundáns módon kibővíteni (csak épp itt a bővítést azáltal érhetjük el, ha aszerint képzünk – itt fölösleges – feltételt, hogy az *Alter* jó volt-e (*G*) vagy rossz (*B*) akkor, amikor az *Ego* cselekvését figyeltük):

$$\begin{aligned}\text{ASR}(\text{IS}, Ego, t_i) \leftrightarrow (\text{G}(Alter, t_i) \rightarrow \mathbf{A}\{\text{DO}(Ego, \text{a}, t_i) \rightarrow \text{VALUE}(Ego, t_i) = \text{VALUE}(a, t_i)\}) \wedge \\ (\text{B}(Alter, t_i) \rightarrow \mathbf{A}\{\text{DO}(Ego, a, t_i) \rightarrow \text{VALUE}(Ego, t_i) = \text{VALUE}(a, t_i)\})\end{aligned}$$

A formula értelmezése: ha *Alter* jó ($\text{G}(Alter, t_i)$), akkor *Ego*-nak olyan $\text{VALUE}(Ego, t_i)$ reputációs minősítést kell kapnia a közösségtől (**A**), aminek értéke megegyezik saját cselekvésének a értékével, és a formula második konjunktív összetevője ugyanezt a tartalmat írja le arra az esetre, amikor az *Alter* rossz ($\text{B}(Alter, t_i)$). A formula (főleg a rövid alakjában) felfedi az *Image Scoring* módszertan néhány fontos vonását. Az új minősítés kiadásakor nincs szükség semmilyen információra a másikkal (*Alter*-rel, *Recipiens*-sel) kapcsolatban, de nem kell figyelembe venni az *Ego* reputációját sem. Ez nagyon egyszerűvé, egyszerűen követhetővé teszi ezt a szabályt. Csak *Ego* cselekvését kell megfigyelni, semmi mást. Ráadásul első pillanatra nagyon is intuitívnak tűnhet, hiszen a kooperációt

jutalmazza, a dezertálást bünteti. Még az a kérdés is feltehető, hogy egyáltalán szükség van-e más szabályra, nem mindig és mindenáron kellene-e a kooperációt ilyen módon, e szabály alkalmazásával bátorítani. Nos, a válasz: nem mindig és nem mindenkinek lehet jó ez a szabály. Ha ezt az egyszerű szabályt írjuk elő, akkor ez arra ösztönözheti az embereket, hogy csak a látszatra figyeljenek, és így a szabály alkalmazása könnyen hipokrízishez, álszentséghez vezethet. Nem véletlen, hogy Jézus a Hegyi beszédben olyan sokat támadja a hipokrita magatartást: „Mikor imádkoztok, akkor se legyetek olyanok, mint a képmutatók, mert ezeknek az kedves, ha a zsinagógákban és a piacok szegletein állva úgy imádkozhatnak, hogy az emberek látják őket. Én azonban azt mondom, elveszítik jutalmukat" (Mt. 6,5). Az *Image Scoring* szabály többet ad a látszatra, mint kéne.

(ii) Az *Image Scoring* módszertannal szemben azt a kérdést lehet feltenni, hogy miért is kéne kooperálni egy rossz (*B*) partnerrel szemben. Ha a rossz személlyel szembeni kooperálást jónak tartjuk, akkor ezzel – lehet, hogy akaratlanul – azt is kifejezzük egyben, hogy nem baj, hogy ő rossz, és mivel nyilván azért lett rossz, mert nem kooperált korábban, az sem baj, ha nem kooperált. Ezzel pedig – még ha csak akaratlanul is, de – a dezertálást bátorítjuk. Éppen ezért mondhatjuk, hogy ha egy rossz partnert büntetünk azzal, hogy nem segítjük őt (vagyis dezertálunk), akkor ezt a büntetést igazoltnak tekinthetjük, és ezért utána jónak (*G*) minősíthetjük az így cselekvő személyt. Ez a megfontolás adja a következő értékelési szabály, a *Simple Standing* (SS) módszertan lényegét (magyarul *egyszerű hírnévápoló* szabálynak nevezhetnénk). Ez a stratégia a jó partnerrel szemben ugyanazt az értékelést tartalmazza, mint az *Image Scoring* szabály, ám a rossz (*B*) *Recipiens*-sel kapcsolatba került *Ego* dezertálását (*D*) jónak (*G*) minősíti. Mindezt a következő módon fejezhetjük ki:

$$\text{ASR}(\text{SS}, Ego, t_i) \leftrightarrow (G(Alter, t_i) \rightarrow \mathbf{A}\{\text{DO}(Ego, a, t_i) \rightarrow \text{VALUE}(Ego, t_i) = \text{VALUE}(a, t_i)\}) \wedge (\text{B}(Alter, t_i) \rightarrow \text{A}\{\text{g}(Ego, t_i)\})$$

A képletből látható, hogy ez a módszertan az *Ego* cselekvése mellett már figyelembe veszi a *Recipiens* reputációját is. Ahogy már jeleztem, ez a szabály az igazolt büntetés ideológiájaként (gyakorlati indokaként) értelmezhető. Amikor a gyerekeinknek meséket mondunk, nagyon sokszor zárjuk azzal a történetet, hogy „a rossz elnyerte méltó büntetését". Történelmi tapasztalataink és az új keletű játékelméleti szimulációk egyaránt az elv igazát bizonyítják, „ha rossz voltál, meg kell büntetni", vagy, ami ugyanaz: meg kell dicsérni azt, aki a rosszat tevőt megbünteti. Erről szól a *Simple Standing* módszere (Ohtsuki – Iwasa 2004, 2006, 2007).

(iii) De ezzel még nincs vége. Az igazolt büntetés intézményének elfogadása után még egyet léphetünk tovább. Lehetünk még szigorúbbak a büntetésben. Ha a rosszat minden formában büntetni akarjuk, mert hiszünk abban, hogy így gátat vethetünk a terjedésének, akkor feltehetjük a kérdést magunknak: „Jó dolog-e kooperálni egy rossz személlyel, szabad-e segítenünk a rosszat?" A *Simple Standing* szabály szerint megengedhető, ám a *Stern Judging* (SJ) -módszertan szerint ez elfogadhatatlan (Pacheco 2006; Brandt – Sigmund 2004). Ezt az új mechanizmust nevezték még *Kandori*-szabálynak

(Kandori 1992), és bár a szakirodalomban sok hivatkozás van e módszertanra az utóbbi néven, én maradok az előbbinél. A *Stern Judging* szabály lényege:

$$\text{ASR}(\text{SJ}, \textit{Ego}, t_i) \leftrightarrow ((\text{G}(\text{Alter}, t_i) \rightarrow \mathbf{A}\{\text{DO}(\textit{Ego}, \text{a}, t_i) \rightarrow \text{VALUE}(\textit{Ego}, t_i) = \text{VALUE}(a, t_i)\}) \wedge (\text{B}(\textit{Alter}, t_i) \rightarrow \mathbf{A}\{\text{DO}(\textit{Ego}, a, t_i) \rightarrow \text{VALUE}(\textit{Ego}, t_i) \neq \text{VALUE}(a, t_i)\}))$$

Jó *Recipiens* esetében megint csak ugyanazok a szabályok érvényesek, mint az előző két esetben. A rossz partnerrel szemben viszont pont az ellenkezőjét kell csinálni annak, mint amit az *Image Scoring* módszertana írt elő. Ez a szabály kérlelhetetlen a rossz emberekkel szemben, ami abban nyilvánul meg, hogy amennyiben a *Donor* kooperál a rossz *Recipiens*-sel szemben, tehát segíti őt, akkor a szabály ezt a segítséget rossz minősítéssel bünteti. Rosszakkal nem lehet együttműködni azáltal, hogy nem büntetjük őket, mert: „vétkesek közt cinkos, aki néma". Egy teljesen hétköznapi példával élve: amikor egy babazsúr után a mama felfedezi, hogy a csokitortát a méregdrága perzsaszőnyegbe pacskolták a gyerekek, és a saját gyerekét kérdőre vonva („Miért tettétek ezt?") azt a választ kapja, hogy „Nem én voltam, hanem a Józsika", akkor az anya azonnal rávágja, „De hát miért nem szóltál rá!?". Ez a *Stern Judging* logikája: büntetni azt, aki nem büntet. Ez a szigorúság nem rontja le a módszertan azon képességét, hogy hatékonyan tudja promotálni a kooperációt.

(iv) A büntetési logikát lehet még tovább szigorítani, de ez már a kooperáció kialakulásának útjában áll. A legerősebben büntető értékelési szabály a *Shunning* (SH) módszertan (olykor nevezik szigorú diszkriminátorszabálynak is). Ez a szabály már három esetben is bünteti a cselekvőt azzal, hogy rossznak minősíti (Ohtsuki – Iwasa 2007; Panchanathan – Boyd 2004; Nowak 2006). A stratégiát leíró formula így néz ki:

$$\text{ASR}(\text{SH}, \textit{Ego}, t_i) \leftrightarrow (\text{G}(\textit{Alter}, t_i) \rightarrow \mathbf{A}\{\text{DO}(\textit{Ego}, a, t_i) \rightarrow \text{VALUE}(\textit{Ego}, t_i) = \text{VALUE}(a, t_i)\}) \_ (\text{B}(\textit{Alter}, t_i) \rightarrow \mathbf{A}\{\text{B}(\textit{Ego}, t_i)\})$$

Bizonyos szempontból hasonlít a *Simple Standing* módszertanra, hiszen ez is feltétel nélküli szabályt ír elő a rossz *Recipiens*-sel szemben, csak épp itt a *Donor* reputációja feltétel nélkül rossz lesz (míg a *Simple Standing* a jó minősítést osztja ki feltétel nélkül). A *Shunning* módszertan azt „üzeni" a cselekvőnek, hogy „tehetsz bármit, ha rosszal kerülsz szembe, te is rosszá válasz". Ezt fejezi ki a mondás: „Aki korpa közé kerül, megeszik a disznók". A játékelméleti szimulációk tanúsága szerint ez az értékelési szabály nem (vagy csak ritkán, különös helyzetekben) támogatja a kooperációt. Nem véletlen: a *Shunning* módszertan a szegregáció logikája. Arra ösztönöz, hogy elkerüljük a rossz(nak minősített) emberek társaságát, hogy kirekesszük őket.

(v) Ahogy megmutattuk a négy DR szabályt egy összefoglaló táblában, úgy most megtesszük ugyanezt az IR-mechanizmusokkal is *(2. táblázat)*.

2. táblázat: *A négy értékelési szabály mintázata*

ASR ↓	VALUE(*Alter*, t_i) →	G		B	
	DO(*Ego*, *a*, t_i) →	C	D	C	D
Image Scoring	**A**{VALUE(*Ego*, t_i)=*x*} ⇉	G	B	G	B
Simple Standing		G	B	G	G
Stern Judging		G	B	B	G
Shunning		G	B	B	B

A négy szabály ugyanúgy értékeli a jó *Recipiens*-sel szembeni cselekvést: a jó partnert támogatni kell, a rosszat büntetni (a támogatás megvonásával). A különbség a rossz *Recipiens*-sel kapcsolatos szabályokban van (Rosas 2010). A büntetés erőssége mentén haladó érvelést már bemutattam, de azt a kérdést még nem tettem fel, hogy egyáltalán miért van szükségünk többféle értékelési rendszerre.

Korábban már említettem az igazolt büntetés intézményét, most újból előveszem ezt a kategóriát. Ez a fogalom ugyanis egy furcsa paradoxon okozója. Hogy ezt belássuk, szükségünk van a John Searle konstitutív szabályát megalapozó 'counts as' szabály felidézésére (Searle 1969). Gondoljunk bele, mit is jelent megbüntetni valakit. Valakivel szemben valami rosszat tenni (vagyis dezertálni), és azt mondani rá, hogy „ez a rosszat tevés most büntetésnek számít". Amikor persze ezt tesszük, akkor egyben ugyanarra a cselevésre vonatkozóan kétféle interpretációs lehetőséget is konstituálunk. Az értékelési szabályaink pedig ezt a kétértelmű – mert kétféle minősítésű – cselekvést kell, hogy megítéljék. Mielőtt továbblépnék, két megjegyzést kell itt tennem. (i) A büntetés mindig egy másik cselekvésre vonatkozik, ezért másodrendű fogalom. Ebből vezethető le az a tény, hogy ha felvesszük a büntetés intézményét a játékelmélet modelljébe, akkor szükségszerűen meg fog jelenni a másodrendű fogolydilemma jelensége is (Oliver 1980; Boyd – Richerson 1992). (ii) A büntetés mint olyan szükségképpen mindig kétféle módon interpretálható, értékelhető. Amikor büntetünk, mindig megteszünk valami rosszat vagy tartózkodunk valami jó megtételétől (például pénzbüntetést szabunk ki vagy nem nyújtunk segítséget). Ilyenkor mindig azt mondjuk, hogy amit ilyenkor cselekszünk, az büntetésnek számít – a searle-i értelemben vett intézményi szinten. De ez – mint minden jóravaló másodrendű fogalom esetében – paradoxonhoz vezet. Nézzük meg, miért.[14]

Ahogy jeleztem, a másodrendű fogolydilemma akkor keletkezik, amikor a játékok világába beemeljük a büntetés intézményét. De miért is alakul ez ki? Azért, mert a büntetésnek mindig van költsége, viszont a büntetés olyan kollektív jószág, aminek a létrejöttéhez, fenntartásához nem kell mindenkinek hozzájárulnia. Elegendő, ha a közösségből csak néhányan büntetik a dezertáló személyeket. Ekkor viszont az lesz a kérdés, hogy ki vállalja magára a büntetés költségeit, és mivel mindenki erős kísértést érezhet arra, hogy a másiktól várja ezt el, ezen az új szinten egy má-

14 Ezt a paradoxont korábban már megírtam máshol (Syi 2009).

sodrendű fogolydilemma alakul ki. Feltételezhetjük, hogy az elsőrendű alapjátékban a kooperálás támogatást, segítséget jelent, míg a dezertálás ennek megtagadását. A másodrendben viszont minden a fordítottja lesz ennek: a kooperációt ott az jelenti, hogy valaki büntet (vagyis megvonja a támogatást), míg a büntetés elhagyása ott dezertálásnak számít. Ha ezt a két szintet egy táblázatban jelenítjük meg, akkor láthatóvá válik az ellentmondás *(3. táblázat)*.

3. táblázat: *A büntetés paradoxona*

	másodrendű kooperálás = büntetni = nem segíteni	másodrendű dezertálás = nem büntetni = segíteni
elsőrendű kooperáció = segíteni	ellentmondás	segíteni
elsőrendű dezertálás = nem segíteni	nem segíteni	ellentmondás

A büntetés paradoxona nyilvánvaló: ha segítünk, akkor kooperálunk az elsőrendű értékelés alapján, de dezertálunk a másodrendű interpretáció szerint, és ugyanezt a kettős értelmezési lehetőséget kapjuk a nem segítés esetén is. Mindig választanunk kell az értelmezési lehetőségek közül, sosem lesz egyértelmű a helyzet. Ezért van szükségünk több értékelési rendszerre. Az *Image Scoring* bünteti a büntetést magát, a *Simple Standing* jutalmazza a büntetést, a *Stern Judging* bünteti a nem büntetést, vagyis jutalmazza a másodrendű büntetést, a *Shunning* pedig bünteti a cselekvőt magát (nem pedig a cselekvést).

(vi) Az eddigi négy értékelési rendszerünk másodrendű volt abban az értelemben, hogy két paramétert használt fel: az *Ego* cselekvését és az *Alter* reputációját. Úgy lehet finomítani az indirekt reciprocitás modelljén, hogy harmadrendű értékelési szabályokat is engedünk felvenni. A harmadrendű szabályok már figyelembe veszik az *Ego* reputációját is. Ebben az esetben elméletileg 4096 stratégia képzelhető el, de ezek közül – a szimulációs számítások szerint – csak nyolc olyan stratégia van, amely kooperációképesnek mondható (Ohtsuki – Iwasa 2006, 2007). Talán még fontosabb az a tény, hogy az ez ebbe a nyolc stabil, kooperációkész stratégiába tartozó nyolc darab harmadrendű értékelési rendszer dekomponálható a már ismert négyféle másodrendű értékelési szabály segítségével. Mindennek belátásához csak annyit kell tennünk, hogy a harmadrendű értékelési rendszereket felbontjuk, aszerint, hogy a jó vagy a rossz *Ego* esetén milyen szabályokat írnak elő *(4. táblázat)*.

4. táblázat: *A nyolc harmadrendű értékelési szabály dekomponálása*

ASR ↓	*Good Ego*	*Bad Ego*
Standing	Simple Standing	Image Scoring
Simple Standing	Simple Standing	Simple Standing
ASR_1	Simple Standing	Stern Judging
Strict Standing	Simple Standing	Shunning
ASR_2	Stern Judging	Image Scoring
ASR_3	Stern Judging	Simple Standing
Stern Judging	Stern Judging	Stern Judging
Judging	Stern Judging	Shunning

Az első oszlop mutatja a harmadrendű szabályok nevét. Van, ami ismerős, van, ami nem. A *Standing* vagy a *Judging* szabályok a *Simple Standing,* illetve a *Stern Judging* módszertanok harmadrendű változatai (Sugden 1986). Látható, hogy van pár szabály, aminek még nevet sem adtak a szakirodalomban (ASR_i). Tudjuk, hogy léteznek, de még nem találtunk megfelelő leírást hozzájuk. Ami azonban igazán fontos, az az, hogy a nyolc harmadrendű szabály felbontható olyan módon, hogy a jó *Ego*-ra vonatkozóan négy-négy esetben a *Simple Standing,* illetve a *Stern Judging* módszertant alkalmazhatjuk csak, és ezután mindkét csoportban „kioszthatjuk" a négy másodrendű szabályt a rossz *Ego* esetében. Mindez azt jelenti, hogy a harmadik logikai szinten nincs szükségünk további szabályok felvételére, nem kell további ontológiai elköteleződéseket tennünk, másodrendű fogalmakkal le tudunk írni mindent. Ez a tény tovább növeli a már bemutatott négyféle másodrendű értékelési szabály jelentőségét, hiszen segítségükkel modellezhetjük az IR világ egész jelenségét.

A kétféle reciprocitás közös mintázata

Ha a kétféle reciprocitáshoz tartozó kooperációképes stratégiák szabályszerűségeit általánosabb formában reprezentáljuk, érdekes közös mintázatokat fedezhetünk fel. A DR cselekvési szabályokat összegző táblázatban *Ego* cselekvéseit írtuk le (a C és D szimbólumok segítségével), az IR értékelési szabályok pedig a reputációs értékelésekből álltak (amiket *G* és *B* szimbólumokkal jeleztünk). Mivel mind a cselekvéseket, mind a reputációs minősítéseket egy fekete-fehér világképen alapuló dichotóm logika szerint értékeljük, a *C*, illetve *D*, valamint a *G*, illetve *B* értékek ugyanazt jelentik. Ezt kifejezhetjük azzal is, hogy 1, illetve 0 értékekkel helyettesítjük őket. Ekkor egyetlen táblázatban foglalhatjuk össze a kétféle reciprocitás szabályait *(5. táblázat).*

5. táblázat: *A négy stratégiapáros közös mintázata*

ACR ↓	C		D		← DO(*Ego*, *a*, t_{i-1})
	C	D	C	D	← DO(*Alter*, a, t_{i-1})
Simple Tit-for-tat	1	0	1	0	Image Scoring
Contrite Tit-for-tat	1	0	1	1	Simple Standing
Pavlov	1	0	0	1	Stern Judging
Grim Tigger	1	0	0	0	Shunning
DO(*Ego*, *a*, t_i) →	C	D	C	D	
VALUE(*Alter*, t_i) →	G		B		↑ ASR

A nyolc DR és IR stratégia között felfedezhetünk négy összetartozó kettőst. A *Simple TFT* és az *Image Scoring*, a *Contrite TFT* és a *Simple Standing*, a *Pavlov* és a *Stern Judging*, valamint a *Grim Trigger* és a *Shunning* szabályok azonos mintázatokat mutatnak (Rosas 2010). Ez az alaki hasonlóság tartalmilag a büntetéshez való viszonyban nyilvánul meg.

Mind az *Image Scoring*, mind a *TFT* elsőrendű szabály, és a reciprocitás dinamikájában betöltött szerepük is hasonló: „Úgy véljük, hogy a Scoring szerepe az indirekt reciprocitásban hasonló, mint a Tit-for-tat szabályé a direkt reciprocitáson belül. Egyik stratégia sem evolucionárisan stabil, de az a képességük, hogy katalizálják a kooperáció kialakulását kedvezőtlen körülmények között, valamint az egyszerűségük, jelentőséget kölcsönöz nekik" (Nowak 2006). Mindkét szabály ellenzi a büntetést – igaz, hipokrita módon. Közös elvük: büntesd, aki büntet, amit mondhatunk úgy is, hogy 'ne büntess'.

Amikor Sugden javasolta a T_1 (*Standing*) stratégiát (Sugden 1986), nem beszélt reputációról, de ahhoz nagyon közeli fogalmat használt: 'standing'. Később a *Standing* (*Simple Standing*) mint értékelési szabály az IR elemzések fontos részévé vált. Aztán amikor Boyd újracímkézte (*Bűnbánó TFT*) Sugden kategóriáját (Boyd 1989), akkor ő már cselekvési szabályként hivatkozott rá. Terminológiazavarnak kell tekintenünk, hogy hol cselekvési, hol értékelési szabályként hivatkoznak erre a stratégiára? Nem. Szerintem inkább a két fogalom közös vonásaira utal ez az „összekeverhetőség". Amikor Frean bevezette a szigorú, de igazságos (*Firm-but-Fair*) stratégia fogalmát, a következő indoklást adta: „…»szigorú«, mert dezertálással torolja meg, ha az előző menetben szívott (ő kooperált, a partnere dezertált). Viszont »igazságos«, mert nem torolja meg a partner dezertálását, ha vele együtt ő is dezertált korábban, és nem folytatja tovább a partnere kizsákmányolását akkor, ha az előző menetben egyszer már kihasználta őt" (Frean 1994). Ez a gondolatmenet nagyon hasonló ahhoz, amit a *Bűnbánó TFT* szabály bemutatásakor vázoltam fel, bár az érvelés hangsúlya „eltolódott" a büntetés irányába. *Ego* úgy értékeli *Alter* előző fordulóbeli dezertálását, mint ami igazolt büntetés volt (*Ego* előző fordulóbeli dezertálásával szemben), és mivel *Ego* egyetért az igazolt büntetés intézményével, és egyben igazságos (nem részrehajló), így *Ego* kooperál. A két stratégia közös jelszava: 'ne büntesd, aki büntet'.

A korai jellemzések hangsúlyozták a *Pavlov* szabály tanulási képességét. Rosas hízelgő, alázatos stratégiaként jellemezte a *Pavlov/Stern Judging* kettőst (Rosas 2010). A két szabály közös vonását a büntetéshez fűződő viszonyulásukban találhatjuk meg. A *Pavlov*

stratégia megegyezik abban a *Stern Judging* szabállyal, hogy *Ego* úgy értékelheti az *Alter* előző fordulóbeli kooperálását, miközben *Ego* ekkor dezertált, hogy *Alter* tartózkodott a büntetéstől. Ez pedig elfogadhatatlan – még akkor is, ha maga az *Ego* volt az, aki dezertált, és akit emiatt büntetni kellett volna. A két stratégia hasonlít egymásra abban, hogy helyteleníti a büntetéstől való tartózkodást. Jelszavuk: 'büntesd, aki nem büntet'.

A haragtartó (*Grim Trigger*) és a *Shunning* stratégiák közös vonása az, hogy könnyen kivezetnek a kooperációból. Ellenzik a cselekvés mindkét formáját, büntetik mind a büntetés végrehajtását, mind a büntetéstől való tartózkodást. A túlzott szigor azonban megöli az együttműködést, könnyen a kapcsolatok teljes felszámolásához vezet. Az elv, ami falat emel közénk: 'büntesd, bármit tesz'.

A négy alapelv tehát, amiből választhatunk, a következő: a 'ne büntess/büntesd a büntetést', a 'ne büntesd a büntetést', a 'büntesd a nem büntetést', valamint a 'büntess mindent'. Az elvek hátterét adó érvelési logikát bemutattam. Lehetne azzal az erkölcsi tanulsággal zárni, hogy rajtunk áll, melyik elvet követjük, és azok vagyunk, amit választunk, de szociológiailag nem ez az igazán fontos. Az persze igaz, hogy sokat elmond a személyiségünkről, hogy éppen melyik elvet tartjuk követendőnek. De egyfelől ezek az elvek inkább a közösségi érzületek jellemzésekor fontosak, másfelől pedig az változhat időben, illetve kontextusról kontextusra, hogy egy adott közösség (személy) éppen melyik elv szerint ítélkezik a tényleges társadalmi gyakorlatában. Valamelyik elv adott pillanatban domináns lehet egy közösségen belül, ám idővel felválthatja azt egy másik. A közösségek történelmét vélhetőleg folytonos hullámzás jellemzi, amely az enyhébb büntetési „politika" felől halad az egyre erősebb felé, majd vissza. A két „végpontot" a *TFT/Image Scoring,* illetve a *Grim Trigger/Shunning* stratégiakettős jelenti. Az előbbi a mindenáron kooperálás, az utóbbi a mindenáron büntetés hipokríziséhez vezethet. Köztük vannak a mérsékelt büntetés elvét követő párok (a *Contrite TFT/Simple-Standing,* illetve a *Pavlov/Stern Judging*), amelyek egyébként a legsikeresebbek a kooperáció hosszú távú fenntartásában. Az a tény, hogy a gyerekeinknek elmondott első mesétől kezdve életünk végéig rengeteg narratívába belefoglaljuk a „rossz elnyeri méltó büntetését" tanulságot (tehát a *Simple Standing* logikát érvényesítjük), azt jelzi, hogy ezt az elvet minden társadalom, minden közösség fontosnak tartotta, tartja. „Békés, nyugodtabb" időszakokban ez az elv elégséges lehet a kooperáció megfelelő szintjének fenntartásához. Amikor mégsem, akkor egyre erősödik a büntetés iránti elvárás, ami a büntetés elmulasztásának számonkérésében testesül meg (ami a *Stern Judging* elve). Amíg ez az elvárás nem széles körű, még nem veszélyezteti a kooperáció fennmaradását, ám minél több emberre érvényesítik, annál inkább és egyre erősebben a szegregációs (*Shunning*) logikát erősítik. Miután pedig kiderül, hogy ez nem oldja meg a kooperációs problémákat, az inga elindul a másik irányba, vagyis a közösség elkezd lazítani a büntetés szigorán. Természetesen az effajta társadalmi dinamika nem annyira a társadalom egészét, mint inkább a kisebb-nagyobb közösségeket, szubkultúrákat jellemzi, és egy adott pillanatban a bemutatott elvek egymásba fonódó, kusza egymás mellett élését tapasztalhatjuk – sokféle mozgás, sokféle irány, sokféle intenzitás mentén. E dinamikák

felderítése érzékeny terepmunkát kívánna meg. Tanulmányomban csak azt akartam megírni, hogy ennek során milyen elveket, milyen mintákat lenne érdemes keresni.

Abstract: Reciprocity can help the evolution of cooperation. We use the concept of strategy to model the two types of reciprocity (direct and indirect). In the case of direct reciprocity there are four second-order action rules (Simple Tit-for-tat, Contrite Tit-for-tat, Pavlov, and Grim Trigger), which are able to promote cooperation. In the case of indirect reciprocity the key component of the cooperation is the assessment rule. There are, again, four elementary second-order assessment rules (Image Scoring, Simple Standing, Stern Judging, and Shunning). The eight concepts can be formalized with the help of only two operators (action and value), two agent concepts, and the notion of goodness. The formalism helps us to discover that the action and assessment rules can be paired, and they show the same patterns. The logic of these patterns can be interpreted with the concept of punishment that has an inherent paradoxical nature.

Irodalom

Alexander, R. D. (1987): *The Biology of Moral Systems*. Aldine de Gruyter.

Axelrod, R. (1984): *The Evolution of Cooperation*. Basic Books.

Belnap, N. – Perloff, M. –Xu, M. (2001): *Facing the Future*. Oxford Univ. Press.

Boerlijst, M. C. – Nowak, M. A. – Sigmund, K. (1997): The Logic of Contrition. *Journal of Theoretical Biology,* 185(3): 281–293.

Boyd, R. (1989): Mistakes Allow Evolutionary Stability in the Repeated Prisoner's Dilemma Game. *Journal of Theoretical Biology,* 136(1): 47–56.

Boyd, R. – Richerson, P. J. (1989): The Evolution of Indirect Reciprocity. *Social Networks*, 11: 213–236.

Boyd, R. – Richerson, P. J. (1992): Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology,* 13(3): 171–195.

Brandt, H. – Sigmund, S. (2004): The Logic of Reprobation: Assessment and Action Rules for Indirect Reciprocation. *Journal of Theoretical Biology*, 231(4): 475-486.

Frean, M. R. (1994): The Prisoner's Dilemma Without Synchrony. *Proceeding of the Royal Society Biological Siences,* 257(1348): 75–79.

Friedman, J. W. (1971): A Non-cooperative Equilibrium for Supergames. *Review of Economic Studies,* 38(1): 1–12.

Gouldner, Alvin W. (1960): The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review,* 25(2): 161–178.

Hauert, Ch. – Schuster, H. G. (1998): Extending the Iterated Prisoner's Dilemma without Synchrony. *Journal of Theoretical Biology,* 192(2): 155–166.

Imhof, L. A. – Fudenberg, D. – Nowak, M. A. (2007): Tit-for-tat or Win-stay, Lose-shift? *Journal of Theoretical Biology,* 247(3): 574–580.

Kandori, M. (1992): Social Norms and Community Enforcement. *The Review of Economic Studies,* 59(1): 63–80.

Kanger, S. – Kanger, H. (1966): Rights and Parliamentarism. *Theoria,* 32(2): 85–115.

Kraines, D. – Kraines, V. (1989): Pavlov and the Prisoner's Dilemma. *Theory and Decision,* 26(1): 47–79.

Kraines, D. – Kraines, V. (1993): Learning to Cooperate with Pavlov an Adaptive Strategy for the Iterated Prisoner's Dilemma with Noise. *Theory and Decision,* 35(2): 107–150.

Malinowski, B. (1972): *Baloma*. Budapest: Gondolat Kiadó.

Mauss, M. (2004): *Szociológia és antropológia*. Budapest: Osiris Kiadó.

Nowak, M. A. – Sigmund, K. (1993): A Strategy of Win-stay Lose-shift that outperforms Tit-for-tat in the Prisoner's Dilemma Game. *Nature*, 364: 56–58.

Nowak, M. A. (2006): Evolutionary Dynamics of Cooperation. *Proceedings of the International Congress of Mathematicia*, 3(3): 1523–1540.

Nowak, M. A. – Sigmund, K. (1998): The Dynamics of Indirect Reciprocity. *Journal of Theoretical Biology,* 194(4): 561–574.

Nowak, M. A. – Sigmund, K. (2005): Evolution of Indirect Reciprocity. *Nature,* 437(7056): 1291–1298.

Ohtsuki, H. – Iwasa, Y. (2004): How Should we Define Goodness? – Reputation Dynamics in Indirect Reciprocity. *Journal of Theoretical Biology,* 231(1): 107–120.

Ohtsuki, H. – Iwasa, Y. (2006): The Leading Eight: Social Norms that can Maintain Cooperation by Indirect Reciprocity. *J. Theor. Biol.,* 239(4): 435–444.

Ohtsuki, H. – Iwasa, Y. (2007): Global Analyses of Evolutionary Dynamics and Exhaustive Search for Social Norms that Maintain Cooperation by Reputation. *Journal of Theoretical Biology,* 244(3): 518–531.

Oliver, P. (1980): Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Ivestigations. *American Journal of Sociology,* 85(6): 1356–1375.

Pacheco, J. M. – Santos, F. C. – Chalub, F. A. C. C. (2006): Stern Judging: a Simple, Successful Norm which Promotes Cooperation under Indirect Reciprocity. *PLoS Computational Biology,* 2(12): 1634–1638.

Panchanathan, K. – Boyd, R. (2003): A Tale of Two Defectors: the Importance of Standing for Evolution of Indirect Reciprocity. *Journal of Theoretical Biology,* 224(1): 115–126.

Panchanathan, K. – Boyd, R. (2004): Indirect Reciprocity can Stabilize Cooperation without the Second-order Free Rider Problem. *Nature,* 432(7016): 499–5024.

Ridley, M. (1997): *The Origins of Virtue.* London: Viking Press.

Rosas, A. (2010): Evolutionary Game Theory Meets Social Science: Is there a Unifying Rule. *Journal of Theoretical Biology,* 264(2): 450–456.

Ruzsa I. (2000): *Bevezetés a modern logikába*. Budapest: Osiris Kiadó.

Sahlins, M. D. (1973): Törzsek. *Vadászok, törzsek, parasztok*. Budapest: Kossuth Kiadó, 135–315.

Searle, J. (1969): *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Sigmund, K. (2010): *The Calculus of Selfishness*. Princeton: Princeton University Press.

Sigmund, K. – Nowak, M. A. (1998): Evolution of Indirect Reciprocity by Image Scoring. *Nature,* 393(6968): 573–577.

Sugden, R. (1986): *The Economics of Rights, Co-operation and Welfare*. Oxford: Basil Blackwell.

Syi (2008): *Cselekvéselmélet dióhéjban*. Budapest: Typotex.

Syi (2009): Önzetlenségelméletek. *BUKSZ,* 21(3): 226–233.

Syi (2012): Feléd fordítom, újra. Bajnok A. – Korpics M. – Milován A. – Pólya T. – Szabó L. (szerk.) *A kommunikatív állapot: Diszciplináris rekonstrukciók*. Budapest: Typotex, 122–134.

Takahashi, N. – Mashima, R. (2003): *The Emergence of Indirect Reciprocity: Is the Standing Strategy the Answer?* Hokkaido Univ. WP no.29. (http://lynx.let.hokudai.ac.jp/COE21/pdf/029.zip)

Trivers, R. (1971): The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46(1): 35–57.

Von Wright, G. H. (1963): *Norm and Action*. London: Routledge and Kegan Paul.