EWOLUCYJNE PODSTAWY INSTYTUCJI SPOŁECZNYCH

Piotr Swistak
University of Maryland

W pracy tej opisuję wynik, który wskazuje na ewolucyjne podstawy instytucji społecznych. Instytucje społeczne uważane są w socjologii za zjawiska typowo strukturalne, a zatem niewytłumaczalne z mikro-poziomu racjonalnie działających jednostek. Okazuje się jednak, że trzy typy instytucji społecznych – sankcje wobec osób trzecich, konformizm i normy zinternalizowane - dają się dedukcyjnie wyprowadzić jako warunki konieczne stabilnej równowagi w grach ewolucyjnych. Artykuł ten w sposób nieformalny opisuje teorię i twierdzenie, które wynik ten ustala[1].

Luka mikro-makro

Jednym z najciekawszych i bodaj najbardziej rozpowszechnionych problemów nauki jest teoretyczna luka pomiędzy teoriami wyjaśniającymi zjawiska na pozio-

Piotr Swistak, Department of Government and Politics, University of Maryland, College Park, Md 20742, e-mail: pswistak@gvpt.umd.edu

mie mikro i teoriami struktur makro. Najsłynniejszy przykład takiej luki istnieje zapewne w fizyce. Teoria kwantowa doskonale działa na poziomie mikro-świata cząstek elementarnych. Podobnie teoria względności jest dobrym opisem świata makro. Głównym problemem fizyki nie jest zatem to, że nie potrafi ona wyjaśnić jakiejś klasy zjawisk, ale to, że nie potrafi to zrobić w ramach jednej wewnętrznie spójnej teorii. Problem ten polega na sformułowaniu teorii, która łączyłaby w sobie obydwa poziomy analizy tzw. teorii wszystkiego (*theory of everything*)[2].

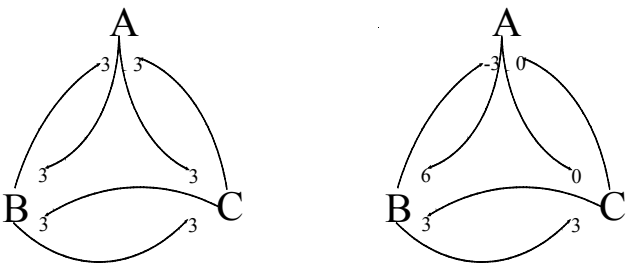
Problem teoretycznej luki między teoriami mikro i makro występuje również w naukach społecznych (Coleman 1986, 1990, Eggertsson 1990). Mikro teorie, na przykład teorie podejmowania decyzji, wyjaśniają zachowania jednostek, makro teorie, jak teoria cykli rynkowych, wyjaśniają zachowania struktur. Brak jest ogólnych teorii, które łączyłyby obydwa poziomy analizy. Jednocześnie wydaje się, że pełne zrozumienie podstawowych struktur, jak grupa czy państwo, nie jest możliwe bez istnienia takich teorii. Indywidualizm metodologiczny w ogólności, a teorie wyboru racjonalnego w szczególności, starają się wywieść własności struktur i instytucji z założeń o jednostkach i o tym jak podejmują one decyzje. Hobbes, na przykład, wykazuje w „Lewiatanie", dlaczego racjonalny człowiek w stanie natury stworzy państwo. Podobnie Rawls pokazuje, dlaczego w interesie racjonalnej jednostki jest zaadaptowanie określonej zasady redystrybucji (tzn. zasady sprawiedliwego podziału) w społeczeństwie.

Czy możliwe są tego rodzaju wyjaśnienia instytucji społecznych? Niektóre dziedziny nauk społecznych opierają się na założeniu, nie zawsze jawnym, że instytucji nie da się wyprowadzić z własności jednostek. Typowa analiza socjologiczna, na przykład studium o samobójstwie Durkheima, przyjmuje istnienie instytucji społecznych i bada ich wpływ na zachowania jednostek. Z jednej strony paradygmat mikro stara się wyprowadzić instytucje z własności jednostek, które je tworzą, z drugiej, paradygmat makro stara się wyprowadzić własności jednostek z instytucji, do których jednostki te należą. Pomiędzy teoriami mikro i teoriami makro pozostaje luka, której istnienie stworzyć może nowe kierunki rozwoju w naukach społecznych. Gdyby w istocie okazało się, że instytucje społeczne dają się wyprowadzić jako własności interakcji pomiędzy racjonalnymi graczami, to duża część teorii socjologicznej wymagałaby radykalnego przemyślenia.

W artykule tym chciałbym naszkicować idee pewnej analizy, która łączy poziom analizy makro z poziomem analizy mikro. Chciałbym również pokazać, jak w analizie tej daje się wydedukować konieczność istnienia podstawowych instytucji społecznych. Aby lepiej wyjaśnić intuicje, które kierować będą moją konstrukcją, zacznę swoje rozważania od następującego przykładu:

Przykład

Rysunek 1



(a) (b)

Rozważmy dla przykładu grupę trzech osób ─ A, B i C, które wchodzą ze sobą we wszystkie możliwe interakcje tzn. A z B, A z C i B z C. Załóżmy, że interakcje te powtarzają się w czasie i każda z nich ma identyczną strukturę. Załóżmy, na przykład, że w każdej interakcji każdy gracz może być wobec drugiego kooperatywny lub też nie. Jeśli obaj będą kooperatywni, to każdy z nich uzyska wypłatę 3. Jeśli jeden gracz jest kooperatywny, a drugi nie, to osoba wykorzystująca kooperatywność drugiej dostaje wypłatę 6, a osoba wykorzystana, wypłatę -3. W przypadku wreszcie, gdy gracze są niekooperatywni, obaj otrzymują wypłatę 0. Czytelnik traktować może wypłaty jak sumy pieniędzy tzn. 3 zł, 6 zł, itd. Ze względów czysto stylistycznych rozróżniać będę pomiędzy wypłatą i użytecznością. Wypłatę proponuję utożsamiać z jej obiektywnym wskaźnikiem np. sumą pieniędzy, użyteczność natomiast z subiektywną wartością wypłaty dla gracza[3].

Rozważmy teraz szereg sytuacji, w których, w standardowych modelach rynku, wypłata dla gracza A jest taka sama. Na rysunku 1a, na przykład, A otrzymuje wypłatę 6 jako rezultat kooperatywnych interakcji z B i C; na rysunku 1b A otrzymuje identyczną wypłatę, tzn. 6, wykorzystując B i nie kooperując z C. Jeśli A jest *homo oeconomicus* to dla A wartość (użyteczność) dwóch różnych zachowań, które generują te same wypłaty, powinna być identyczna. Inaczej mówiąc, dla *homo oeconomicus* użyteczność wypłaty zależy jedynie od jej wysokości, a nie od tego, „jak" ta wypłata została uzyskana. Przez *homo sociologicus* rozumieć będę gracza, dla którego warunek ten nie jest spełniony. Dla *homo sociologicus* użyteczność wypłaty nie jest jedynie funkcją jej wysokości. Dowolny czynnik, który ma wpływ na użyteczność wypłaty dla gracza, nazywać będę instytucją. Jedyną instytucją ekonomiczną jest zatem sama wypłata, jako że dla *homo oeconomicus*

użyteczność wypłaty jest jedynie (rosnącą) funkcją wypłaty. Wszystkie inne czynniki, które mogą mieć wpływ na użyteczność, nazywać będę instytucjami społecznymi.

Patrząc na rysunek 1 nietrudno jest sobie wyobrazić jak różne instytucje społeczne mogą wpływać na użyteczność graczy. Rozważmy, na przykład, gracza A, który wyznaje moralną zasadę niewykorzystywania innych. Jego użyteczność w sytuacji *a* (lewej) będzie oczywiście wyższa, *ceteris paribus*, od jego użyteczności w sytuacji *b* (prawej). Podobnie, jeśli A jest egalitarystą[4] jego użyteczność w sytuacji 1a będzie wyższa niż w sytuacji 1b. Tak samo będzie w przypadku, gdy A jest konformistą, a zatem osobą, dla której źródłem dodatkowej użyteczności jest fakt, że jej zachowanie jest identyczne z zachowaniem większości graczy w grupie. Podobnie inne instytucje społeczne, jak na przykład zazdrość, snobizm, nonkonformizm, dyskryminacja, normy zinternalizowane, mogą mieć również wpływ na użyteczności graczy.

Jeśli instytucja społeczna prowadzi do istotnych odstępstw od użyteczności *homo oeconomicus,* to będzie ona zmieniać zachowania *homo oeconomicus*. Zmiany zachowań, natomiast, prowadzić będą do zmian w rozkładzie wypłat. Na przykład egalitarysta, który nie toleruje dużych różnic w wypłatach, może podjąć działania obniżające wypłatę swoją jak również wypłaty innych, jeśli zmniejszy w ten sposób nierówności w grupie. Podobnie, jeśli norma B przeciwko niewykorzystywaniu innych jest wystarczająco silna, B może przestać kooperować z A, aby ukarać A za wykorzystywanie C (w sytuacji 1b). Takie postępowanie może zakończyć kooperacje z A i, w konsekwencji, obniżyć wypłatę B. Oznacza to, że zachowanie, które maksymalizuje użyteczność gracza, prowadzić może do obniżenia jego wypłaty. Ponieważ wszystkie instytucje społeczne stanowią, z definicji tego terminu, ograniczenia na zachowanie maksymalizujące wypłatę, wszystkie one są z natury rzeczy w konflikcie z zasadą maksymalizacji wypłaty. Warto przy tym zauważyć, że nasz *homo sociologicus,* mimo że nie jest graczem maksymalizującym wypłatę, jest graczem maksymalizującym użyteczność. A zatem jest to gracz racjonalny w standardowym rozumieniu pojęcia racjonalności.

Jeśli naszym celem jest wyjaśnienie pojawienia się instytucji społecznych w Hobbsowskim świecie, który jest ich pozbawiony, nasz problem postawić można w sposób następujący: W jaki sposób świat pełen maksymalizujących wypłaty *homo oeconomicus* zmienia się w świat *homo sociologicus*? Lub inaczej, w jaki sposób rynek zamienia się w grupę społeczną? Jak jedna forma zachowania racjonalnego (jeden typ funkcji użyteczności, ten, który zależy tylko od wypłat) zamienia się w inny? Jakie są źródła preferencji społecznych? A zatem, jak możemy na poziomie makro wyjaśnić pojawienie się instytucji, które nie maksymalizują wypłat członków grupy?

Natura rozwiązania

Odpowiedź na nasze pytania wymaga sformułowania teorii działania racjonalnego, która byłaby na tyle ogólna, aby mogła opisać nie tylko *homo oeconomicus* (tj. standardowe modele teorii gier), ale również *homo sociologicus* z potencjalnie nieskończoną różnorodnością definiujących go instytucji społecznych. Załóżmy przez chwilę, że udało nam się sformułować taką teorię. Załóżmy, co więcej, że ma ona postać gry. Dobrze zdefiniowane są zatem trzy pojęcia teorii: gracze, strategie i wypłaty (użyteczności). Jeśli w takiej teorii chcemy wyjaśnić istnienie określonej instytucji społecznej, to musimy udowodnić, że instytucja ta jest koniecznym warunkiem rozwiązania (punktu równowagi) tej gry[5]. A zatem krytyczna część takiej analizy sprowadza się w gruncie rzeczy do założeń, pod którymi szukać będziemy punktów równowagi. Jak przy każdej analizie dedukcyjnej, natura założeń determinować będzie zawartość i znaczenie rozwiązania. Dokładne zrozumienie pojęć i założeń teorii jest w tym sensie równoważne ze zrozumieniem rozwiązania.

Założenia, na których opiera się opisana tu analiza, wynikają z następujących dwóch rozważań ogólnych. Po pierwsze, ponieważ moim celem jest wyjaśnienie powstania instytucji społecznych, swoją analizę muszę zacząć od stanu, w którym instytucje te nie istnieją. A zatem punktem wyjściowym takiej analizy jest „stan natury", w którym istnieje tylko *homo oeconomicus*; jej celem jest pokazanie, że w punkcie równowagi takiej gry pojawić się muszą normy społeczne. Drugim bardzo istotnym warunkiem analizy jest to, aby rozwiązanie (tzn. punkty równowagi gry) nie opierało się o żadne silne założenia. Mówiąc inaczej, moim drugim celem jest uzyskanie rozwiązania pod bardzo ogólnymi założeniami. Ponieważ mam nadzieję wykazać, że pewne formy instytucji społecznych są zarazem bardzo powszechne i bardzo trwałe, wydaje się być oczywiste, że wynik taki oparty być musi na odpowiednio ogólnych założeniach. Nie miałby on sensu, gdyby okazało się, że zależy w sposób istotny od „stopnia racjonalności" graczy, dokładnego sprecyzowania wypłat, statycznej formy gry, określonej dynamiki zmian zachowań graczy, spełnienia założeń gry przez wszystkich dokładnie graczy (tj. wykluczenie dewiantów), czy też innych warunków, które nie byłyby odporne na drobne perturbacje w specyfikacji gry[6].

Podstawowym wynikiem, który tłumaczy pojawienie się instytucji społecznych, jest twierdzenie orzekające, że stabilny rynek interakcji nie istnieje bez instytucji społecznych (Bendor i Swistak 1998). Mówiąc dokładniej, w grach, w których interakcje mają postać wymiany kooperatywnej, instytucje społeczne są koniecznym warunkiem zachowań stabilnych (tj. zachowań w punkcie równowagi gry). A zatem grupa społeczna okazuje się być rynkiem wymiany, który wytworzył mechanizmy stabilizujące. Trzy różne instytucje społeczne, normy spo-

łeczne, konformizm i normy zinternalizowane okazują się być konieczne do stabilizacji zachowań *homo oeconomicus* (wszystkie trzy pojęcia zostaną dokładniej zdefiniowane później). Poprzez połączenie mikro i makroanalizy zachowań ludzkich, teoria ta rzuca nowe światło na problem powstania i problem stabilności norm społecznych.

Teoria, w której daje się udowodnić to twierdzenie, jest teorią dedukcyjną. W artykule tym postaram się naszkicować jej główne pojęcia, założenia i wyniki. Będzie to, z założenia, opis nieformalny. Pełna i formalnie poprawna konstrukcja została przedstawiona gdzie indziej (Swistak 2003)[7].

Forma interakcji międzyosobowych

Jeśli interakcja między dwoma graczami pozwala każdemu z nich na uzyskanie maksymalnej wypłaty niezależnie od tego, co zrobi drugi gracz, to ich wybory są strategicznie trywialne: ponieważ to, co jest najlepsze dla każdego gracza, może on uzyskać niezależnie od innych, nie ma w tej sytuacji żadnego konfliktu interesów, a zatem nie istnieje powód, aby powstała jakakolwiek instytucja, która zawsze działa jako ograniczenie na zachowania. Grę, która posiada tę własność nazywać będę, za Bendorem i Swistakiem (1998), trywialną. A zatem, żeby zrozumieć ewolucję instytucji musimy zacząć naszą analizę od interakcji, w których możliwy jest konflikt, to znaczy od gier, które nie są grami trywialnymi. Aby lepiej opisać znaczenie wprowadzanych pojęć i sens przyjmowanych założeń, proponuję ograniczyć dalszą dyskusję do dobrze znanej i prostej do interpretowania klasy gier.

Rozważmy zatem najprostszy teoriogrowy model konfliktu (jednokrotną grę postaci 2 x 2), który pojawia się w następującym problemie wymiany rynkowej pomiędzy A i B. Powiedzmy, że A posiada chleb, który B chętnie by od niego kupił płacąc 2 złote za bochenek. Załóżmy ponadto, że A byłby skłonny sprzedać chleb za tę cenę. A zatem obydwaj gracze zyskaliby na wymianie. Mimo to do wymiany może nie dojść. Jeśli zachowania graczy nie są regulowane żadnymi prawami (stan natury), najlepszym wyjściem dla A jest ukraść pieniądze B i zatrzymać chleb. Podobnie najlepszym wyjściem dla B jest zatrzymać pieniądze i ukraść chleb. Drugim dopiero w kolejności preferencji (obydwu graczy) rozwiązaniem jest wymiana, trzecim, brak wymiany, a najgorszą ze wszystkich możliwych jest sytuacja, w której gracz traci swoją własność nie uzyskując w zamian nic. Problem „wymienić czy ukraść" ma strukturę tzw. dylematu więźnia (DW). Ta sama struktura wypłat występuje w wielu innych typach interakcji np. w interakcjach towarzyskich. Rozważmy, na przykład, problem wymiany przysług. Z punktu widzenia egoisty najlepszą jest sytuacja, w której inni świadczą mu

przysługi, a on ich nie odwzajemnia (świadczenie przysług jest kosztowne), a najgorszą sytuacją, w której on wyświadczył przysługę (poniósł koszt) nie dostając nic w zamian; jednocześnie obydwaj gracze preferują wymianę przysług nad jej brak. Logika problemu „wymienić czy ukraść" i problemu wymiany przysług jest, jak się okazuje, identyczna. Fakt ten stanie się jasny po zdefiniowaniu szczegółów interakcji w formie gry.

Rozważmy dwóch graczy, A i B, którzy wybrać muszą między dwoma możliwościami: kooperacją i defekcją. Gracz A może zatem kooperować z B (np. wymienić towar, wyświadczyć przysługę) lub też nie (defekcja). Wybory B są identyczne. Załóżmy dalej, że A uzyskuje największą wypłatę w sytuacji, gdy A wybierze defekcję, podczas gdy B będzie kooperować. Oznaczmy użyteczność tej sytuacji dla A przez T i załóżmy, dla ustalenia uwagi, że T=5. Kolejną najbardziej preferowaną sytuacją dla A jest sytuacja obopólnej kooperacji. Powiedzmy, że A uzyskuje w tym przypadku wypłatę R=3. Jeśli obaj gracze wybiorą defekcję, A uzyskuje wypłatę P=1. Wreszcie najniższą wypłatę, S=0, A uzyskuje w sytuacji, gdy jego kooperacja spotka się z defekcją B. Załóżmy wreszcie, że porządek wypłat jest taki sam dla B, a jego wypłaty (użyteczności) są odpowiednio T*, R*, P* i S*. Ponieważ wypłaty obydwu graczy spełniają warunki T>R>P>S oraz T*>R*>P*>S*, gra ta jest przykładem tzw. dylematu więźnia. Defekcja w tej grze jest strategią dominującą: zapewnia ona wypłatę wyższą niż kooperacja niezależnie od tego, co zrobi gracz drugi. A zatem obopólna defekcja jest jedynym punktem równowagi tej gry. W punkcie równowagi gracze (A i B) otrzymują wypłatę (P,P*), która jest gorsza dla obydwu, niż wypłata uzyskana przy obopólnej kooperacji (R,R*).

DW obrazuje problem konfliktu i kooperacji w najostrzejszej formie. Jest wiele innych typów gier, w których problem ten pojawia się również i to w różnym stopniu nasilenia. Wyniki tu opisane stosują się do wszystkich gier nietrywialnych. Ponieważ jednak DW jest najbardziej znanym i najlepiej rozumianym modelem konfliktu, będę go dalej używać jako przykładu w kolejnych krokach analizy.

Gry powtarzalne

Jeśli DW grany jest tylko raz (oraz gracze wiedzą o tym i są racjonalni), wówczas wynikiem gry jest Pareto nieoptymalny wektor wypłat – obaj gracze mogliby uzyskać wypłaty wyższe, gdyby kooperowali. Załóżmy teraz, że interakcja nie ogranicza się do jednokrotnego spotkania. Gracze spodziewają się zatem przyszłych interakcji, być może wielu, nie wiedząc nigdy czy kolejna interakcja będzie ich ostatnim spotkaniem, czy też nie. Formalnie sytuację taką możemy mo-

delować zakładając, że gracze wchodzą w nieskończenie wiele interakcji, z których każda ma postać jednokrotnego DW[8]. Strategia (czysta) gracza A w grze z graczem B zdefiniowana jest jako kompletny plan gry. A zatem jest to funkcja, która określa ruch gracza w każdej iteracji k w zależności od historii gry (ruchów obydwu graczy) we wszystkich iteracjach poprzednich tj. od 1 do k-1. Strategia jest zatem teoriogrowym modelem normy zachowania. Takie pojęcie strategii jest oczywiście znacznie bardziej wymagające niż potoczne pojęcie normy zachowania. Strategia w teorii gier jest kompletnym planem gry, niezależnie od tego, jaki ruch wykona przeciwnik, strategia określa nam dokładnie, co mamy robić w każdej sytuacji. Tak wymagające pojęcie jest konieczne ze względów analitycznych. Bez niego nie moglibyśmy określić wypłat w grze w sposób jednoznaczny. Gdyby sekwencja ruchów między dwoma normami zachowań nie była określona jednoznacznie, nie bylibyśmy w stanie sprecyzować wypłat w takiej grze. (Później uogólnimy pojęcie strategii na normy, które nie będą musiały być pełnymi planami gry.)

A zatem pojęcie strategii wyznacza prosty sposób oceniania zachowań. Kiedy dwóch graczy, z których każdy ma kompletny, deterministyczny plan gry (strategię czystą), wchodzi ze sobą w interakcje, ich strategie wyznaczają jednoznacznie wszystkie ruchy w grze. Ponieważ w każdej iteracji graczom zależy na wypłatach w iteracjach przyszłych, rozsądnym wskaźnikiem wartości danej strategii przy ustalonej strategii przeciwnika jest, na przykład, średnia wypłata na iterację[9]. Na przykład, kiedy obydwaj gracze grają strategię ZAWSZE DEFEKCJA (ZD) – graj defekcję bezwarunkowo we wszystkich iteracjach gry – każdy gracz otrzymuje wypłatę P=1 w każdej iteracji, a zatem średnia wypłata na iterację uzyskiwana przez strategię ZD w grze z ZD jest 1. Rozważmy teraz sytuację, w której zamiast ZD jeden z graczy przyjmuje strategię ALT – kooperuj w iteracjach nieparzystych i defektuj w parzystych. ALT grając przeciw strategii ZD uzyska ciąg wypłat 0,1,0,1,... (jako że S=0 oraz P=1), a zatem średnią wypłatę na iterację równą 0,5. Oczywiście, średnia wypłata na iterację będzie sensownym wskaźnikiem wartości strategii tylko wtedy, gdy graczowi zależy na wypłatach uzyskanych w przyszłych iteracjach. Moja analiza ogranicza się do takich właśnie gier[10].

Punkty równowagi w grach iterowanych

Aby zidentyfikować punkty równowagi w grach iterowanych, w których przyszłe wypłaty są wystarczająco ważne, rozważmy dla przykładu strategię WET ZA WET (WZW): kooperuj w pierwszej iteracji gry, a później w dowolnej iteracji k rób to, co twój oponent zrobił w iteracji k-1. Załóżmy, że gracz A używa strategii WZW. Jak powinien grać B, który znając strategię A, chce zmaksymalizować

swoją wypłatę? Jeśli przyszłe wypłaty są dla B wystarczająco ważne, to powinien on kooperować z WZW we wszystkich iteracjach gry. Defekcja w dowolnej iteracji może jedynie obniżyć jego wypłatę (zakładając, że P<R oraz R>(T+S)/2). A zatem B powinien grać strategię, która kooperuje we wszystkich iteracjach z WZW. Jedną z (nieskończenie) wielu takich strategii jest WZW. (Każda inna strategia może co najwyżej uzyskać tą samą wypłatę co WZW.) Załóżmy zatem, że B gra WZW i załóżmy dodatkowo, że A wie, że B gra strategię WZW. Ale w tej sytuacji rozumowanie, które zastosowaliśmy powyżej do gracza B, zastosować możemy teraz do gracza A. WZW jest najlepszą odpowiedzią na WZW: żaden z graczy, wiedząc, że drugi gra WZW, nie może zwiększyć swojej wypłaty grając strategię inną. Strategie te są zatem w równowadze. W języku teorii gier mówimy, że para strategii WZW w iterowanym DW, w którym przyszłe wypłaty są dla graczy wystarczająco ważne, jest w równowadze Nasha. Nie jest to jednak jedyna para strategii w równowadze Nasha. Jeśli, na przykład, A wierzy, że B gra strategię ZD, to A powinien grać defekcję w każdej iteracji. Strategia ZD jest przykładem strategii, która będzie najlepszą odpowiedzią A na strategię ZD gracza B. Ale to samo rozumowanie zastosować możemy do B. Jeśli B wie, że A gra ZD, to jego najlepszą odpowiedzią jest ZD. Para strategii ZD jest zatem również w równowadze Nasha. W ogólności okazuje się, że istnieje nieskończenie wiele różnych strategii w równowadze Nasha, które generują dowolną ilość[11] kooperacji od 0% (para strategii ZW) do 100% (para strategii WZW). A zatem, jeśli gra ograniczona jest do dwóch graczy, standardowe pojęcie rozwiązania przewiduje, że w równowadze (Nasha) pojawić się może nieskończenie wiele różnych norm zachowań. Czy taką nieskończoną różnorodność zachowań da się również zaobserwować w równowadze, gdy rozszerzymy naszą teorię do wielu graczy (wchodzących w iterakcje parami)?

Homo oeconomicus

Na rynkach, rozważanych w formie idealnej, jedynym celem graczy jest maksymalizacja wypłaty. A zatem decydując pomiędzy dwiema strategiami, *homo oeconomicus* preferuje strategię, która daje mu większą wypłatę i jest indyferentny pomiędzy strategiami, które dają mu wypłaty identyczne[12].

Drugie oblicze *homo oeconomicus* dotyczy struktury społecznej interakcji. Standardowo zakłada się (zazwyczaj w sposób niejawny), że strategia *homo oeconomicus* A w grze z B będzie taka sama niezależnie od tego, czy A i B są jedynymi graczami, którzy wchodzą ze sobą w interakcje, czy są oni członkami większej grupy[13], czy też są członkami różnych grup. Inaczej mówiąc, jeśli A jest *homo oeconomicus,* to jego zachowania wobec B będą jedynie zależeć od zachowań B wobec A, nie będą one natomiast zależeć od żadnych interakcji z osobami trzeci-

mi. Warunki te definiują, w sposób nieformalny, pojęcie *homo oeconomicus* i pojęcie rynku jako nietrywialnej gry w zbiorze *homo oeconomicus*.

Aby lepiej zrozumieć własności *homo oeconomicus,* dobrze będzie przyjrzeć się wybranym implikacjom tego pojęcia. Załóżmy, na przykład, że B jest znany z tego, że gra strategię WZW, gracz A natomiast rozważa, czy użyć WZW, czy też BO (BEZWZGLĘDNY ODWET: kooperuj w iteracji pierwszej i nigdy nie graj defekcji tak długo, jak przeciwnik kooperuje; jeśli przeciwnik wybierze defekcję w jakiejkolwiek iteracji k, graj defekcję we wszystkich iteracjach począwszy od k+1). Ponieważ WZW i BO generują identyczne zachowania – wzajemną kooperację we wszystkich iteracjach gry – wypłaty gracza A są w obu przypadkach identyczne. Jako *homo oeconomicus* A będzie zatem indyferentny pomiędzy WZW i BO. Jego użyteczność obu strategii jest taka sama. Mówiąc ogólniej, użyteczność strategii zależy tylko od wypłaty, którą strategia ta przynosi, a nie od tego, jak ta wypłata została uzyskana. Również, jeśli A jest *homo oeconomicus* i jeśli dwie różne strategie przynoszą A tę samą wypłatę w grze z B, to użyteczność tych dwóch strategii powinna być dla A identyczna we wszystkich grupach, w których A i B wchodzą w interakcje. Mówiąc inaczej, zachowania *homo oeconomicus* są niezależne od struktury społecznej interakcji.

Pojęcie *homo oeconomicus* jest jednym z najbardziej użytecznych analitycznie narzędzi w naukach społecznych. Wydaje się jednocześnie, że jest to pojęcie deskryptywnie puste – w rzeczywistości *homo oeconomicus* nie istnieje. Ponieważ zatem *homo sapiens* nie zachowuje się zgodnie z modelem *homo oeconomicus*, podstawowym problemem każdej teorii działania racjonalnego jest wyjaśnienie, dlaczego gracze maksymalizujący użyteczność działają inaczej niż gracze maksymalizujący wypłatę.

Aby wyjaśnić, jak *homo oeconomicus* przekształcił się w *homo sociologicus,* musimy stworzyć system pojęć, w którym możliwe będzie opisanie zarówno rynku (świat bez instytucji społecznych), jak i grupy (świat z instytucjami społecznymi). Gdybyśmy aksjomatycznie przyjęli model interakcji, w którym instytucje społeczne nie mogą się pojawić, nie moglibyśmy w takim modelu wyjaśnić ich powstania.

Homo sociologicus

Odchylenia od modelu *homo oeconomicus* stanowią poniekąd podstawowy aspekt ludzkich zachowań. Ludzie karzą i nagradzają innych nie tylko za to, co zrobili w stosunku do nich, ale również za to, co zrobili w stosunku do innych członków ich grupy (normy społeczne). Nasze użyteczności często zależą od roz-

kładu wypłat w grupie, nie tylko od naszej wypłaty. Często cenimy bardziej wypłatę uzyskaną w grupie z mniejszymi nierównościami wypłat (egalitarianizm). Czasami staramy się powiększyć wypłaty najbiedniejszych członków grupy nawet kosztem zmniejszenia wypłaty własnej (rawlsizm). Często cenimy bardziej wypłatę uzyskaną w grupie z większą wypłatą średnią, nawet jeśli nasza własna wypłata jest w niej mniejsza (utylitarianizm). W przypadkach ekstremalnych nasza użyteczność może być całkowicie niezależna od wypłaty własnej i być jedynie funkcją wypłat otrzymywanych przez innych (np. ekstremalny altruizm). Często zdarza się też tak, że nasza użyteczność jest tym większa, im mniejsze są wypłaty innych (zazdrość). Typowo czerpiemy większą użyteczność zachowując się zgodnie z zachowaniami większości członków grupy (konformizm), aczkolwiek czasami sprawiają nam większą przyjemność zachowania nonkonformistyczne. Często norma zachowania ma dla nas wartość sama w sobie niezależnie od wypłaty, którą generuje (norma zinternalizowana). Jeśli, na przykład, cenimy sobie normę wybaczania, czerpać będziemy większą użyteczność grając wybaczającą strategię WZW, niż nie wybaczającą strategię BO, mimo że obydwie strategie dają nam tę samą wypłatę[14].

Te i inne odchylenia od modelu *homo oeconomicus* mieszczą się w dwóch kategoriach zachowań: (1) gracze uzależniają działania w stosunku do siebie od innych interakcji w grupie, i (2) użyteczność strategii nie jest jedynie funkcją wypłaty. Aby wyjaśnić pojawienie się tych odchyleń, musimy wprowadzić pojęcia, które pozwolą nam na zdefiniowanie tych odchyleń. Pierwsza kategoria odchyleń wymaga uogólnionego pojęcia strategii, druga uogólnionego pojęcia funkcji użyteczności.

Aby wyjaśnić ewolucję *homo oeconomicus,* musimy wprowadzić siatkę pojęć, która pozwoliłaby na to, aby instytucje społeczne pojawiły się w świecie *homo oeconomicus*. Musimy to zrobić tak, aby instytucje te nie stanowiły sztywnych ograniczeń – instytucje społeczne muszą być analitycznie możliwe nie będąc jednocześnie analitycznie konieczne. Moim celem jest wyjaśnienie pojawienia się takich instytucji w świecie, który początkowo pozbawiony jest tego rodzaju ograniczeń strukturalnych. A zatem, nasza analiza opierać się musi na teorii, w której możliwi są zarówno *homo oeconomicus*, jak i *homo sociologicus*. Tylko, kiedy analitycznie dopuścimy możliwość pojawienia się wszystkich form zachowania, będziemy mogli zrozumieć, dlaczego niektóre z nich powstały, a inne nie.

Uogólnione pojęcie strategii

Homo oeconomicus A uzależnia swoje przyszłe zachowanie w stosunku do B tylko od historii interakcji między tymi graczami. Strategie BO i WZW są przy-

kładami takich zachowań. Jeśli chcemy zrozumieć, dlaczego A może chcieć uzależnić swoje postępowanie wobec B od tego, jak B zachował się w stosunku do innego gracza C, pojęcie strategii musi być wystarczająco ogólne, aby zawarło w sobie tego rodzaju zachowania. Uogólnione pojęcie strategii dopuszczać zatem musi zachowania, w których zachowanie A wobec B zależy nie tylko od tego, co zaszło między A i B, ale również od tego, co zaszło między B i innymi członkami grupy. W ogólności, aby pozwolić na wszystkie możliwe formy warunkowania zachowań, zakładać będę, że strategia A w grze z B pozwala A uzależniać swoje zachowanie od historii wszystkich interakcji, które zaszły między wszystkimi parami graczy w grupie[15].

Uogólnione pojęcie użyteczności

Jeśli chcemy wyjaśnić, na przykład, pojawienie się norm zinternalizowanych, to musimy mieć analityczną możliwość zdefiniowania funkcji użyteczności nie tylko jako funkcji wypłaty, ale również jako funkcji czynników pozaekonomicznych. Jeśli każdy czynnik nieekonomiczny nazwiemy czynnikiem społecznym, to bez utraty ogólności założyć możemy, że użyteczność normy η dana jest funkcją użyteczności u:

$$U(\eta) = u(\text{wypłata}(\eta)) + u(\text{społeczna wartość}(\eta)).$$

Na jednym ekstremum zbioru możliwych funkcji użyteczności znajduje się *homo oeconomicus,* dla którego użyteczność jest jedynie (rosnącą) funkcją wypłaty. Radykalny egalitarysta, którego użyteczność ma wartość 1 wtedy i tylko wtedy, gdy wypłaty wszystkich graczy są identyczne oraz wartość 0 w sytuacji przeciwnej, jest natomiast przykładem ekstremalnego typu *homo sociologicus*.

Własności rozwiązania

Gra ewolucyjna

Wyobraźmy sobie grupę, w której każdych dwóch graczy wchodzi ze sobą w interakcje. Załóżmy, że interakcje w grupie nie mają żadnej struktury – w dowolnej chwili każdy gracz ma takie samo prawdopodobieństwo wejścia w interakcję z każdym innym członkiem grupy. Załóżmy również, że każda interakcja ma ustaloną formę, na przykład jednokrotnego DW, dla każdego gracza wypłaty przyszłe mają istotną wartość i w każdym momencie gry nie jest określony jej koniec. W tej sytuacji sensownym wskaźnikiem wartości interakcji między dwo-

ma graczami jest, na przykład, średnia wypłata na interakcję (jak było to dyskutowane poprzednio). Jeśli grupa składa się z n osób, to każda osoba grać będzie n-1 gier iterowanych. Każda z tych gier generuje wypłatę, której wskaźnikiem, dla ustalenia uwagi, jest średnia wypłata na iterację. Zgodnie z definicją *homo oeconomicus*, jego wypłata w całej grze równa jest sumie wypłat we wszystkich n-1 grach, w których on uczestniczy. Załóżmy teraz, że w momencie rozpoczęcia gry gracze posiadają określone normy zachowania (strategie), ale z upływem czasu, obserwując normy innych, skłonni są swoje zachowania zmieniać. Jest to zatem system dynamiczny, w którym proces obserwacji i uczenia się wywoływać będzie ewolucyjną presję na normy. Presja ta, z kolei, powodować będzie zmianę zachowań: gracze będą zmieniać normy na takie, które dawać im będą większą użyteczność. W wyniku tego procesu normy o większej użyteczności zastąpią normy o użyteczności mniejszej. Znając początkowy rozkład strategii w grupie możemy zatem badać, czy system się ustabilizuje, które normy zanikną, które staną się powszechne itd. Problem stabilności, na przykład, wymaga znalezienia takiej normy, która jeśli zaadaptowana przez wszystkich graczy w populacji i odeprzeć może inwazję dowolnych norm mutantów. Taką normę nazywać będziemy ewolucyjnie stabilną. W naszej teorii normy ewolucyjnie stabilne określają pojęcie rozwiązania gry tzn. punktów równowagi gry.

Zmiana ewolucyjna

Natura zmiany ewolucyjnej jest prosta: im lepsza jest dana strategia w bieżącym pokoleniu, tym większa będzie jej proporcja w pokoleniu następnym. Proces ewolucyjny, inaczej mówiąc, jest procesem dynamicznym, który jest rosnącą funkcją użyteczności: adaptacja polega nam tym, że gracze zmieniają strategie na takie, które dają im większą użyteczność. Możemy oczywiście skonstruować wiele różnych procesów ewolucyjnych, które spełniać będą ten warunek. W biologii formę procesu wyznacza jednoznacznie mechanizm genetyczny rządzący reprodukcją (patrz np. Dawkins 1989). Proces ten nazywa się dynamiką replikacji lub też proporcjonalną regułą adaptacji. W przeciwieństwie do biologii, w naukach społecznych nie istnieje mechanizm, który generowałby specyficzną formę procesu. Kiedy i jak gracze zmieniają swoje zachowanie, zależy od ich funkcji użyteczności, od czynników psychologicznych i społecznych[16], od „stopnia racjonalności" graczy, od tego, co zakładają oni o racjonalności innych i od wielu innych czynników, które w sposób świadomy lub nie wpływają na zachowanie. W tej sytuacji zakładanie, że proces zmiany dany jest przez specyficzną funkcję jest, co najmniej, ryzykowne. Określenie, jak strategie się zmieniają i jakie równania opisują ten proces, może być przedsięwzięciem empirycznie niemożliwym do zrealizowania. A zatem jedynymi sensownymi rozwiązaniami (punktami równowagi)

gry mogą być tylko takie zachowania, które są stabilne pod wszystkimi procesami ewolucyjnymi. Szukając punktów równowagi wymagać będziemy, aby spełniały one ten warunek.

Przekonania o innych

Różni gracze mają różne funkcje użyteczności. A zatem analiza, która w sposób istotny opiera się na założeniu o identyczności wszystkich funkcji użyteczności, jest mało sensowna. Ta sama uwaga dotyczy też innych istotnych parametrów gry, jak na przykład założeń, które gracze czynią o innych. W grach nietrywialnych najlepsza strategia każdego gracza zależy od strategii graczy innych, a te z kolei zależą od ich funkcji użyteczności. Problem gracza polega na tym, że nie zna on funkcji użyteczności innych. Co więcej, nie może on nigdy z pewnością wiedzieć, jakie są te funkcje. Może on, co najwyżej, mieć niesklasyfikowane hipotezy, co do pewnych własności tych funkcji użyteczności. W modelu rynku idealnego zakładamy, że wszyscy gracze są *homo oeconomicus* i wszyscy o tym wiedzą. W modelu świata bardziej realnego, w którym gracze nie muszą być *homo oeconomicus*, założenie to musimy w sposób istotny uogólnić.

Biorąc pod uwagę ogólną formę funkcji użyteczności, zasadne wydają się następujące dwa warunki. Po pierwsze, sensownym wydaje się utrzymanie założenia, że funkcje użyteczności są niemalejącymi funkcjami wypłaty, *ceteris paribus*. Po drugie, ponieważ gracze maksymalizują swoje użyteczności w punkcie równowagi użyteczność społecznej wartości normy musi być dodatnia. Inaczej mówiąc, jeśli w punkcie równowagi gracze utrzymują normę społeczną, to ma ona dla nich pozytywną użyteczność. Gdyby tak bowiem nie było, to gracz zmieniłby się w czystego *homo oeconomicus*[17]. I na tym kończą się założenia co do funkcji użyteczności. W szczególności nie będę zakładać nic, co do relatywnego wpływu wypłaty i społecznej wartości na użyteczność normy. A zatem zakładam, że możliwe tutaj są zarówno dwa ekstrema, tzn. sytuacje, w których użyteczność wypłaty jest równa zero (ekstremalny przypadek *homo sociologicus*) i użyteczność wartości społecznej jest równa zero (*homo oeconomicus*), jak również wszystkie możliwe inne funkcje wypłaty i wartości społecznej. Przekonania graczy co do funkcji użyteczności innych mogą być zatem prawie dowolne. Podobną dowolność chcemy utrzymać w stosunku do założeń na temat przekonań graczy o rozkładzie różnych funkcji użyteczności w grupie. Chcemy zatem dopuścić pełną heterogeniczność przekonań: dowolny rozkład funkcji użyteczności w grupie stanowić może przekonanie któregoś z graczy, że taki właśnie rozkład w grupie tej występuje (zobacz Bendor i Swistak 2000).

Normy i strategie

Używanie strategii jako modelu zachowań ludzkich nie zawsze jest zabiegiem sensownym. Zakładanie, że w grach iterowanych ludzie posługują się kompletnymi planami gry, jest w większości przypadków absurdalne. Z wyjątkiem bardzo prostych strategii jak ZAWSZE KOOPERUJ czy WET ZA WET ludzie rzadko mają wyczerpujące plany gry. W szczególności zakładanie, że plany te mogą być dowolnie złożone jest nonsensowne (np. Simon i Schaeffer 1992). Ludzkie zachowania kierowane są zazwyczaj bardzo prostymi i niekompletnymi regułami. Na przykład, w iterowanym DW gracz może kierować się proskryptywną regułą „nigdy nie graj defekcji pierwszy" lub preskryptywną regułą „zawsze karaj defekcję defekcją". Realna wydaje się również świętobliwa norma bezwarunkowej kooperacji we wszystkich iteracjach gry czy też przesądna norma zabraniająca grania defekcji w iteracji trzynastej[18]. Wszystkie te normy zachowania stanowią określone ograniczenia na możliwe strategie gracza. W ogólności normę zdefiniujemy jako dowolne ograniczenie na zbiorze strategii gracza[19]. Z powyższych uwag wynika, że normy, a nie strategie, powinny być przedmiotem analizy formalnej. Kiedy dalej rozważać będę ewolucję i stabilność zachowań, szukać będę punktów równowagi gry złożonych z norm, a nie tylko ze strategii.

Punkty równowagi (rozwiązania)

Rozważmy grupę osób, w której wszyscy grają tę samą normę. Jeśli norma ta jest najlepszą odpowiedzią na siebie samą (tzn. jest ze sobą w równowadze Nasha), to nikt w grupie nie ma powodu od normy takiej odejść i zachowania pozostaną w stanie równowagi. Ale taka równowaga może być bardzo niestabilna. Jeśli paru graczy zmieni, z jakiegoś powodu, swoje zachowania, stara norma może nie być najlepszą odpowiedzią na nową ekologię zachowań i gracze mogą zacząć od tej normy odchodzić. Bardziej sensowna definicja równowagi powinna wymagać, aby powszechna w grupie norma pozostała w tej sytuacji stabilna. Chcemy zatem zdefiniować pojecie normy stabilnej tak, aby pojawienie się małej ilości mutantów normy tej nie destabilizowało. Norma taka nazywa się ewolucyjnie stabilna. Mówiąc dokładniej, norma jest ewolucyjnie stabilna, jeśli wystarczająco duża częstość występowania tej normy w populacji gwarantuje, że nie zostanie ona zdestabilizowana przez żadną inwazję normy-mutantów[20]. Wszystkie opisane poniżej wyniki, które dotyczą stabilności norm, używać będą to właśnie pojęcie stabilności.

Racjonalne postawy instytucji społecznych

Po nieco bliższym przyjrzeniu się założeniom i podstawowym pojęciom tej analizy, nasza siatka pojęciowa jest już wystarczająco bogata, abyśmy mogli powoli przejść do prezentacji wyników. Aby uniknąć frustracji, ważne jest jednak pamiętać, że naszkicowana tu teoria jest teorią dedukcyjną, a wyniki, o których będzie mowa, to twierdzenia, które wymagają dowodów formalnych. Niektóre dowody są bardziej skomplikowane, inne mniej, wszystkie są jednak na tyle trudne, że zwykła intuicja nie wystarcza, aby zobaczyć, dlaczego wynikają one z naszych założeń. W istocie, jak zobaczymy później, niektóre z nich są zupełnie nieintuicyjne.

Nie istnieje stabilny *homo economicus*

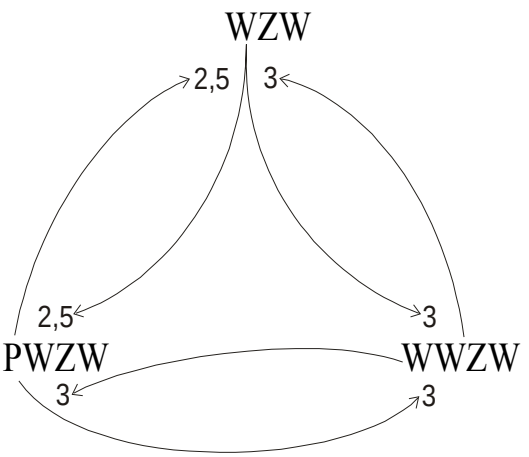
Problem związany z *homo oeconomicus* daje się opisać bardzo prosto: w grach nietrywialnych nie istnieją punkty równowagi składające się z *homo oeconomicus* (Bendor i Swistak 1998). Rozważmy, na przykład, trzech graczy, A, B i C, którzy wchodzą ze sobą w interakcje postaci iterowanego DW. Załóżmy, że podczas gdy początkowo wszyscy kooperują ze wszystkimi, w pewnym momencie A zaczyna grać defekcję w stosunku do B, B natomiast lojalnie kooperuje z A. Załóżmy, co więcej, że podczas gdy B jest wykorzystywany przez A, A i C oraz B i C kooperują ze sobą przez cały czas. W takiej sytuacji „moralny" problem gracza C polega na tym, czy powinien on ukarać A za wykorzystywanie B, czy też nie? Norma nakazująca, aby wrogowie naszych przyjaciół byli naszymi wrogami, wymaga, aby C karał A defekcją. Ale jeśli C zacznie karać A, to ich wzajemna kooperacja z A, która jest korzystna dla C, może się zakończyć. Postępowanie zgodne z normą „wróg mojego przyjaciela jest moim wrogiem" zmniejszy wypłatę C – a to, z punktu widzenia *homo oeconomicus*, nie jest postępowaniem racjonalnym (maksymalizującym użyteczność).

Powyższy przykład opisuje sedno mechanizmu, który destabilizuje maksymalizującego wypłatę *homo oeconomicus*. Szczegóły tego mechanizmu pokazuje następny przykład (Boyd i Lorberbaum 1987), który umieszcza powyższy argument w kontekście iterowanego DW. Załóżmy, że w grupie wytworzyła się norma WZW, która używana jest przez wszystkich graczy. Ponieważ wszyscy grają strategię, która jest przyjazna, a zatem taką, która nigdy nie gra defekcji pierwsza (Axelrod 1984, str. 33), każdy gracz będzie kooperować z każdym we wszystkich iteracjach gry. W szczególności, w interakcjach tych nie będzie wykorzystywana „odwetowość" WZW – gotowość karania defekcji defekcją. Ponieważ własność, która nie jest wykorzystywana, może ulec zanikowi, niektóre strategie WZW mogą się zamienić w bardziej wybaczającą wersję WZW, na przykład WWZW: ko-

operuj w iteracjach 1 i 2, a później w dowolnej iteracji k>2 rób to, co oponent zrobił w iteracji k-1. Załóżmy zatem, że mała grupa mutantów WWZW pojawi się w populacji. Ponieważ WWZW jest, podobnie jak WZW, przyjacielski, wszyscy gracze będą dalej kooperować ze wszystkimi i mutacja ta będzie nierozpoznawalna. Zakładając, że wypłata przy wzajemnej kooperacji jest 3, średnia wypłata na iterację w grze pomiędzy WWZW i WZW będzie również 3.

Rysunek 2

Struktura wypłat w grze pomiędzy trzema strategiami:
WET ZA WET (WZW), PODEJRZLIWY WET ZA WET (PWZW)
i WYBACZAJĄCY WET ZA WET (WWZW)



Załóżmy teraz, że w populacji pojawia się nowa mutacja, PODEJRZLIWY WZW (PWZW), który gra defekcję w iteracji 1, a w dowolnej iteracji k>1 gra to samo, co oponent grał w iteracji k-1. WZW zareaguje na PWZW defekcją w iteracji 2 rozpoczynając wendetę, która spowoduje, że PWZW grać będzie defekcję w iteracjach nieparzystych i kooperację w parzystych przeciwko defekcjom w iteracjach parzystych i kooperacjom w iteracjach nieparzystych ze strony WZW. Konsekwencją tych zachowań będzie oscylujący ciąg wypłat 5,0,5,0,... (dla PWZW) i 0,5,0,5,... (dla WZW) przy założeniu, że wypłaty w jednokrotnym DW są T=5 oraz S=0. W tej sytuacji zarówno PWZW jak i WZW uzyskają wypłatę 2,5. Wendeta nie jest jednak najlepszą strategią przeciwko PWZW. Najlepszą odpowiedzią na PWZW jest zignorowanie jego początkowej defekcji i kontynuowanie kooperacji w iteracji drugiej – a zatem dokładnie to, co robi WWZW. W istocie, w iterowanej grze pomiędzy WWZW i PWZW obydwie strategie uzyskają kooperatywną wypłatę 3. Zauważmy również, że PWZW w iterakcjach z PWZW grać będzie defekcję w każdej iteracji i uzyska w tej grze wypłatę 1.

Rozważmy teraz grupę, w której strategia WZW jest prawie uniwersalna i występuje z częstością 1-$\in$. Nisza o bliskiej zeru częstości, składa się natomiast z $\in_1$ strategii WWZW i $\in_2$ strategii PWZW ($\in_1+\in_2=\in$). Zauważmy teraz, że jeśli grupa składa się z *homo oeconomicus,* to użyteczność WZW równa jest $3(1-\in)+2{,}5\in_2+3\in_1$, użyteczność WWZW jest $3(1-\in)+3\in_2+3\in_1$, a użyteczność PWZW, $2{,}5(1-\in)+\in_2+3\in_1$. Ponieważ WWZW kooperuje ze wszystkimi strategiami, jego wypłata jest największa. Ale to oznacza, że gracze WZW będą chcieli zmienić swoje strategie na takie, które jak WWZW maksymalizowałyby ich użyteczność[21]. Logika tej konstrukcji opisuje mechanizm, który destabilizuje *homo oeconomicus*. Ale jeśli instytucje ekonomiczne (funkcje użyteczności *homo oeconomicus*) wykazują brak stabilności, to proces ewolucyjnej selekcji wyłoni instytucje, przy których równowaga jest możliwa. Z naszych rozważań wynika, że każdy mechanizm stabilizujący stanowić musi odejście od modelu *homo oeconomicus*. A zatem każdy taki mechanizm musi być instytucją społeczną. Nasz problem sprowadza się teraz do zidentyfikowania wśród nieskończonej różnorodności możliwych instytucji tych, które indukować będą stabilność.

Sankcje wobec osób trzecich

Jedną z najbardziej kłopotliwych zagadek paradygmatu zachowań racjonalnych są normy grupowe, które nakazują nam karać i nagradzać innych nie za to, co zrobili w stosunku do nas, ale za to, co zrobili w stosunku do innych członków naszej grupy. W rozważanym poprzednio przykładzie z trzema graczami A, B i C (gdzie B był wykorzystywany przez A, a zatem A był wrogiem B, podczas gdy A i C oraz B i C pozostali przyjaciółmi kooperując ze sobą), norma nakazująca, by wrogowie naszych przyjaciół byli naszymi wrogami, wymaga, żeby C grał defekcję w stosunku do A. Inna norma, która orzeka, że „przyjaciel twojego wroga jest twoim wrogiem", wymagałaby od osób używających WZW, aby grały defekcję z WWZW, który kooperuje z wrogiem WZW[22], PWZW.

Rozważania te sugerują jedno z możliwych rozwiązań problemu niestabilności *homo oeconomicus*. Gdyby gracze WZW zmodyfikowali swoją strategię dodając warunek sankcji wobec osób trzecich wymagający defekcji w stosunku do przyjaciół swoich wrogów, wypłaty w tak zmodyfikowanym przykładzie będą istotnie różne niż poprzednio: użyteczności nowej wersji WZW, WWZW i PWZW będą odpowiednio $3(1-\in)+2{,}5\in_2+\in_1$, $(1-\in)+3\in_2+3\in_1$ oraz $2{,}5(1-\in)+\in_2+3\in_1$. W tej sytuacji, nowa wersja WZW ma największą użyteczność – zakładając oczywiście, że $\in$ jest wystarczająco małe – a to sugeruje, że będzie ona w tej ekologii stabilna.

W istocie okazuje się, że intuicje zawarte w tym przykładzie są bardzo ogólne: sankcje wobec osób trzecich są jedną z instytucji społecznych, które wystarczają, aby norma posiadająca tą własność była ewolucyjnie stabilna (Bendor i Swistak 2001, Swistak 2003). Jeśli norma taka ma wystarczająco wysoką częstość w populacji, to odpowiednio skonstruowany system sankcji wobec osób trzecich jest mechanizmem wymuszającym konformizm dewiantów wobec większości – każda dewiacja od normy większościowej daje mutantowi mniejszą użyteczność niż norma większościowa. Taki mechanizm wymuszający konformizm okazuje się być wystarczający, aby ustabilizować normę. W klasie gier kooperacji, które są uogólnieniem DW, do uzyskania stabilności wystarczają sankcje oparte na następujących dwóch powszechnie występujących normach: „wróg mojego przyjaciela jest moim wrogiem" oraz „przyjaciel mojego wroga jest moim wrogiem" (Bendor i Swistak 2001). Fakt, że istnienie tych norm możemy uzasadnić dedukcyjnie pod bardzo ogólnymi założeniami, tłumaczy, jak się wydaje, powszechność sankcji wobec osób trzecich w systemach społecznych, ekonomicznych i politycznych.

Karanie dewiacji od normy, która używa sankcji wobec osób trzecich, wymaga, aby gracz A wiedział nie tylko o tym, co się dzieje w interakcjach pomiędzy nim i B (to założenie wydaje się naturalnie spełnione), ale również o tym, co dzieje się w interakcjach pomiędzy B i C, jak i każdą inną parą graczy w grupie. To drugie założenie wydaje się, przynajmniej w stosunku do niektórych grup, bardzo silne. Czy zatem normy używające sankcji wobec osób trzecich są jedyną formą instytucji społecznej, która stabilizuje zachowania? Wiemy już, że jest to forma wystarczająca. Czy jest ona również konieczna?

Konformizm

Okazuje się, że istnieją trzy typy instytucji społecznych, które stabilizują normy w grach nietrywialnych. Jedną z tych instytucji, jak opisałem to powyżej, są normy oparte o sankcje wobec osób trzecich. Aby zrozumieć istotę drugiego typu instytucji stabilizujących, wyobraźmy sobie następującą sytuację. Załóżmy, że gracz, który posługuje się określoną normą odczuwa wzrost użyteczności zawsze, kiedy jego oponent posługuje się tą samą normą. Fakt, że dzieli on tę samą normę z kimś innym, stanowi dla niego dodatkowe źródło użyteczności. Tego rodzaju mechanizm definiuje w naszej siatce pojęć powszechnie występujące w grupach zjawisko konformizmu. Udowodnić można (Swistak 2003), że dowolnie mały wzrost użyteczności wystarcza, aby ustabilizować strategię taką jak WZW. Mechanizmem stabilizującym zachowania jest tu „dodatnia poprawka" do funkcji użyteczności *homo oeconomicus*, która definiuje w naszej teorii jedną z najbardziej podstawowych instytucji społecznych ─ konformizm.

Normy zinternalizowane

Aby zrozumieć naturę trzeciego typu instytucji stabilizujących, proponuję powrócić do rozważań na temat niezbędnych informacji, które gracze muszą posiadać, aby instytucje te mogły w grupie zaistnieć. Sankcje wobec osób trzecich wymagają, aby gracze znali wyniki wszystkich interakcji w grupie. Konformizm unika tak silnego założenia, wymagając jednak, aby gracze rozpoznawali, czy przeciwnik posługuje się tą samą normą, co oni. Czy stabilność daje się uzyskać bez tego warunku, czy też konformizm, obok sankcji wobec osób trzecich, jest jedyną formą instytucji społecznych, które stabilizować mogą strategie takie jak WZW. Okazje się, że istnieje instytucja społeczna, która stabilizuje normy przy braku jakiejkolwiek informacji na temat norm innych graczy. Jest to mechanizm internalizacji. Definiuje go w naszej teorii następująca konstrukcja. Wyobraźmy sobie gracza posiadającego określoną normę, którego funkcja użyteczności jest taka, że we wszystkich grach, w których używa tej normy, jego użyteczność wzrasta o pewien czynnik dodatni niezależnie od normy przeciwnika. (Równoważnie założyć możemy, że odstąpienie od normy powoduje spadek użyteczności.) Jest to, oczywiście, efekt występujący przy normach zinternalizowanych. Jak się okazuje, dowolnie mały przyrost użyteczności wywołany przez internalizację wystarcza, aby stabilizować normy (Swistak 2003).

Twierdzenie

I tak opisując kolejno pojęcia, założenia i intuicje związane z ich dedukcyjnymi konsekwencjami dotarliśmy do momentu, w którym suma przekazanych informacji pozwala mi sformułować twierdzenie (Swistak 2003), na którym oparty jest ten artykuł. Twierdzenie to orzeka, że trzy instytucje społeczne – sankcje wobec osób trzecich, konformizm i internalizacja – są jedynymi niezależnymi mechanizmami, które są zarówno konieczne jak i wystarczające, aby ustabilizować zachowania w grupie. Mówiąc dokładniej, sankcje wobec osób trzecich są zarówno konieczne, jak i wystarczające, aby ustabilizować zachowania w sytuacji, kiedy w grupie nie ma ani konformizmu, ani norm zinternalizowanych. Konformizm jest zarówno konieczny jak i wystarczający w sytuacji, kiedy w grupie nie ma sankcji wobec osób trzecich i norm zinternalizowanych. I wreszcie, internalizacja jest mechanizmem zarówno koniecznym jak i wystarczającym w sytuacji, kiedy w grupie nie ma sankcji wobec osób trzecich i konformizmu. A zatem, wszystkie instytucje, które stabilizują zachowania, muszą zawierać przynajmniej jedną z tych trzech podstawowych instytucji społecznych: sankcje wobec osób trzecich, konformizm i internalizację. W tym sensie te trzy elementy organizacji społecznej stanowią podstawę wszystkich zachowań stabilnych.

Sformułowanie powyższego twierdzenia było w pewnym sensie celem mojego artykułu. Ponieważ twierdzenie to opiera się na rozbudowanej siatce pojęć i założeń, wytłumaczenie tego wyniku możliwe było dopiero po ekstensywnym opisie teorii, w której twierdzenie to jest sformułowane. W artykule tym skupiłem uwagę na ogólnym typie instytucji społecznych, które pojawić się muszą na rynkach wymiany jako mechanizmy stabilizujące. Celowo nie stawiałem tu pytań innych. Nie oznacza to, oczywiście, że inne pytania nie są ważne. Na przykład, pytania o podstawowe własności punktów równowagi należą, przynajmniej z punktu widzenia polityki społecznej, do problemów najważniejszych. Czy normy, które pojawią się w punktach równowagi, będą normami kooperatywnymi, czy też nie? Jeśli zarówno normy kooperatywne jak i niekooperatywne zaistnieć mogą w punkcie równowagi, to na czym polegać będzie różnica między takimi stanami równowagi? Pytania te, skądinąd bardzo istotne, są jednak drugorzędne wobec mojego głównego zamierzenia. Celem mojego artykułu było rzucenie światła na mikro-podstawy instytucji społecznych. Głównym wnioskiem tej analizy jest to, że struktura społeczna pojawia się w sieciach interakcji jako zbiór mechanizmów stabilizujących normy zachowań. Jeśli daje się analitycznie wykazać, że trzy podstawowe instytucje społeczne, o których orzeka powyższe twierdzenie, są produktem ewolucji, to być może realny jest program analitycznego wyjaśnienia wszystkich instytucji społecznych na tej samej drodze.

Przypisy

1. Artykuł ten jest nieformalną wersją artykułu Piotra Swistaka „The Emergence and Stability of Social Capital", APSA Meetings, Philadelphia, August 28-31, 2003. Narracja wzięta jest w dużej części z rozdziału „The Rational Foundations of Social Institutions" autorstwa Jonathana Bendora i Piotra Swistaka w książce „Politics from Anarchy to Democracy", pod redakcją I. Morrisa, J. Oppenheimera i K.E. Sołtana, Stanford University Press, Stanford CA, 2004.

2. Od kilkunastu lat rozwija się w fizyce tzw. teoria superstrun, która łączy w sobie teorię kwantową i teorię grawitacji. Jeśli rozwiązane zostaną pewne problemy techniczne, to teoria superstrun może stać się ową teorią wszystkiego, której fizyka poszukuje od początków XX wieku.

3. Jest to rozróżnienie nieortodoksyjne i potencjalnie mylące. Warto dodać, że w teorii gier użyteczność jest terminem pierwotnym i pojęcie wypłaty jest tym samym, co pojęcie użyteczności. Mój zabieg rozdzielenia tych pojęć jest czysto stylistyczny, pozwoli mi on na lepsze wyjaśnienie opisanych dalej koncepcji. Nie jest to jednak zabieg istotny teoretycznie. Całą opisaną tu teorię daje się skonstruować w standardowym języku teorii gier tzn. przyjmując użyteczność za jedyny termin pierwotny.

4. Przez egalitarystę rozumiem gracza, dla którego użyteczność wypłaty jest tym większa im mniejsza jest wariancja wypłat w grupie. Dla skrajnego egalitarysty użyteczność wypłaty będzie jedynie funkcją wariancji wypłat w grupie – im niższa wariancja, tym wyższa uży-

teczność. Użyteczność skrajnego egalitarysty nie jest natomiast funkcją wielkości wypłaty (patrz Lissowski i Swistak 1998).

5. Przy rozwiązywaniu gry (tzn. wyprowadzaniu jej punktów równowagi) okazać się może, że dana instytucja społeczna jest lub nie jest konieczną własnością punktu równowagi. Jeśli instytucja społeczna istnieć może jedynie poza punktem równowagi, to zaniknie ona, kiedy ewolucja zachowań doprowadzi system do równowagi. Pytanie, czy dana instytucja jest własnością punktu równowagi, nie jest problemem czysto teoretycznym. Odpowiedź może mieć ważne implikacje dla polityki społecznej. Jeśli, na przykład, dyskryminacja pojawić się może tylko poza punktami równowagi gry, to najlepszym rozwiązaniem problemu dyskryminacji może być powstrzymanie się od jakiejkolwiek regulacji. Jeśli w istocie dyskryminacja jest samokorygującą się dewiacją rynku (gry), to ewolucyjna dynamika zachowań, gdy gra podążać będzie w kierunku punktu równowagi, sama przyniesie rozwiązanie tego problemu – w punkcie równowagi dyskryminacja przestanie istnieć. Jakiekolwiek próby regulowania mogą ten proces spowolnić lub wręcz odwrócić. Jeśli jednak okaże się, że dyskryminacja możliwa jest w punkcie równowagi, to regulacja może być jedynym rozwiązaniem problemu.

6. Załóżmy, że naszym celem jest wyjaśnienie pojawienia się konformizmu. Powiedzmy, że konformizm definiujemy jako instytucję społeczną, przy której użyteczność danego zachowania jest tym większa, im większa część grupy zachowuje się w ten sam sposób. Trudno byłoby zaakceptować wyjaśnienie konformizmu, które opierałoby się na silnych założeniach. Intuicyjnie jest jasne, że normy takie jak konformizm są powszechne dokładnie dlatego, że są one odporne na małe ─ a być może nawet i duże ─ zmiany w specyfikacji parametrów gry. Żadne rozsądne wyjaśnienie norm tak powszechnych i ogólnych jak konformizm nie może się opierać w sposób krytyczny na założeniach, które nie są wystarczająco ogólne.

7. Teoria ta najbliższa jest tzw. ewolucyjnej teorii gier. Ewolucyjna teoria gier wywodzi się z biologii, a dokładniej z prac Maynarda Smitha i Price (1973) i Maynarda Smitha (1982). Paradygmat ten rozprzestrzenił się bardzo szybko w biologii (np. Hines 1987, Axelrod i Dion 1988, Vincent i Brown 1988) i nieco później w teorii gier, ekonomii i innych naukach społecznych (niektóre przeglądowe artykuły to Friedman 1991, Selten 1991, Mailath 1992, Samuelson 1993; wybrane książki to Fudenberg i Levine 1998, Samuelson 1998, Vega-Redondo 1996, Weibull 1995 i Young 1998) Wczesne prace Roberta Axelroda, w szczególności jego książka „The Evolution of Cooperation" , były niezwykle istotne w popularyzacji tego paradygmatu w naukach społecznych. W moich pracach staram się uogólnić modele ewolucyjne tak, aby dało się w ich podstawowej strukturze wyjaśnić instytucje (społeczne).

8. Na pierwszy rzut oka założenie to wydawać się może nierealne lub wręcz absurdalne (np. Hechter 1992). Przy uważniejszej analizie okazuje się, że gry nieskończenie powtarzalne są rozsądnym modelem interakcji skończenie powtarzalnych, w których gracze nie znają końca gry. Wyjaśnienie tego faktu można znaleźć np. u Rubinsteina (1991).

9. Mówienie o wypłacie dla określonej strategii ma sens tylko wtedy, gdy ustalona jest strategia, z którą ona gra. Warto również dodać, że istnieją inne metody określania wypłat w grach iterowanych (patrz np. Fudenberg i Tirole 1991).

10. Powodem, dla którego sensowne jest ograniczenie analizy do tej klasy gier, jest to, że problem kooperacji poza tą klasą jest trywialny. Punkt równowagi w iterowanym DW, w którym gracze istotnie dyskontują przyszłe wypłaty, jest taki sam jak w jednokrotnym DW – obustronna defekcja. Jeśli przyszłe wypłaty mają niewielką wartość, problem maksymali-

zacji wypłat w grze iterowanej sprowadza się do maksymalizacji wypłat w bieżącej iteracji gry, a zatem strategia ZD jest najlepszą odpowiedzią na każdą strategię w grze iterowanej. Ewolucja i stabilność zachowań kooperatywnych, jak i instytucji, które utrzymywać je mogą w równowadze, możliwa jest tylko w grach, w których wypłaty w iteracjach przyszłych są wystarczająco ważne.

11. Przez ilość kooperacji rozumiem tutaj częstość iteracji, w których obydwie strategie kooperowały. Jest to nieortodoksyjne sformułowanie tzw. twierdzenia potocznego, znanego wyniku w teorii gier (patrz np. Fudenberg i Tirole 1991).

12. Jak już pisałem, w teorii gier pojęcie użyteczności jest pojęciem pierwotnym, które nie jest definiowalne za pomocą innych pojęć teorii. Również moją teorię daje się skonstruować w oparciu o jedno niedefiniowalne pojęcie użyteczności tzn. bez konieczności pojęciowego rozdzielania pomiędzy wypłatami i użytecznościami. Wypłatę definiuje się jako użyteczność uzyskaną przez gracza w pewnych podgrach gier bardziej złożonych. Dokładniej mówiąc, wypłatę gracza A w grze z graczem B definiuje się jako użyteczność, którą A uzyskałby w grze z B, gdyby A i B byli jedynymi członkami grupy. Zauważmy, że jeśli A gra z B tą samą grę, ale w większej grupie, kontekst innych interakcji wprowadza możliwość pojawienia efektów społecznych na użyteczność. Użyteczność, którą A uzyskuje ze skądinąd identycznych interakcji z B, może być zatem inna, jeśli iterakcja ta ma miejsce w grupie niż w sytuacji, w której A i B są jedynymi graczami.

13. Dla potrzeb tej dyskusji zdefiniujmy grupę jako zbiór graczy, w którym każda para graczy wchodzi ze sobą w interakcje.

14. Strategia WZW ma ważną własność bycia wybaczającą: jeśli oponent zacznie kooperować, to jego kooperacja zostanie zawsze odwzajemniona. Strategia BO jest maksymalnie niewybaczająca: pojedyncza defekcja karana jest zawsze bezwarunkową defekcją we wszystkich kolejnych iteracjach gry. Ponieważ w grze z WZW zarówno BO jak i WZW generują identyczną wypłatę, dla gracza, który ceni normę wybaczania WZW może mieć większą użyteczność mimo to, że norma wybaczania nigdy nie będzie realizowana w tych interakcjach.

15. Zauważmy, że takie pojęcie strategii nie wymaga aby gracz uzależniał swoje zachowanie od wyniku innych interakcji w grupie; pojęcie to dopuszcza jedynie jednak taką możliwość.

16. Różne procesy uczenia, jak imitacja lub socjalizacja, wpływać będą na szybkość tej zmiany (Axelrod 1984; Gale, Binmore, and Samuelson 1995; Boyd and Richerson 1985; Cabrales 1993).

17. Mówiąc dokładniej, gracze zakładają, że jeśli norma występuje w punkcie równowagi, to nie może ona zmniejszać użyteczności. Jeśli jednak norma nie występuje w punkcie równowagi, to o jej wpływie na użyteczność graczy nic nie można założyć.

18. Zauważmy, że niektóre normy, jak norma przesądna, ograniczają zachowania gracza w niewielkim tylko stopniu, podczas gdy inne normy, jak norma pełnej kooperacji, są tak wymagające, że wyznaczają w sposób jednoznaczny zachowania gracza we wszystkich iteracjach gry. Ten ostatni typ normy jest strategią w sensie teoriogrowym.

19. North (1990 str. 3) definiuje instytucje jako „...reguły gry w społeczeństwie, lub bardziej formalnie, (...) stworzone przez ludzi ograniczenia, które kształtują interakcje" (tłumaczenie własne). Nasza definicja normy pozostaje w tym samym duchu. Formalnie definiujemy normę jako dowolny zbiór strategii (Bendor i Swistak 2000).

20. Odporność na inwazję rozumieć można na dwa sposoby: mocniejszy, jeśli wymagać będziemy, że normy mutantów będą zanikać pod procesem ewolucyjnym, oraz słabszym, jeśli wymagać tylko będziemy, że ich proporcja nie będzie wzrastać. Ponieważ w grach iterowanych norma rodzima może jedynie powstrzymać wzrost mutacji, jedyną możliwą formą stabilności jest stabilność w sensie słabszym (np. Selten 1983; van Damme 1987). Mówiąc bardziej precyzyjnie, normę nazywać będziemy stabilną, jeśli przy wystarczająco dużej proporcji w grupie (ze skończoną liczbą strategii), proporcja ta nigdy nie będzie maleć. Inni nazywali takie strategie semistabilne (Selten 1983), neutralnie stabilne (Sobel 1993) lub neutralne ESS (Warneryd 1993). Jeśli strategia jest stabilna w tym słabszym sensie, to pewne normy mutantów mogą generować identyczne wypłaty jak strategia stabilna, co znaczy, że pozostawać one będą w grupie bezterminowo.

21. Strategia maksymalizująca użyteczność zależeć będzie od przekonań gracza o strategiach używanych przez innych i od tego, jak szybko te strategie będą się zmieniać. Ponieważ nasze pojęcie stabilności wymaga, aby stabilność została zachowana niezależnie od przekonań graczy i niezależnie od szybkości zmian ich strategii, przykład ten pokazuje, że WZW jest strategią niestabilną.

22. Przez wroga gracza X rozumiemy tu gracza Y, który gra w stosunku do X defekcję z niezerową częstością.

Bibliografia

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert, i Douglas Dion. 1988. *The Further Evolution of Cooperation*. „Science" 242 (December 9): 1385-90.

Bendor, Jonathan, i Piotr Swistak. 1997. *The Evolutionary Stability of Cooperation*. „American Political Science Review" 91: 290-307. Tłumaczenie tego artykułu pod tytułem *Ewolucyjna stabilność kooperacji* pojawiło się w „Studiach Socjologicznych", 1998, 3 (150): 127-171.

Bendor, Jonathan, i Piotr Swistak. 1998. *Evolutionary Equilibria: Characterization Theorems and Their Implications.* „Theory and Decision" 45: 99-159.

Bendor, Jonathan, i Piotr Swistak. 2000. *The Impossibility of Pure Homo Economicus*. Working Paper.

Bendor, Jonathan, i Piotr Swistak. 2001. *The Evolution of Norms*. „American Journal of Sociology" 106: 1493-1545.

Bendor Jonathan, i Piotr Swistak, 2004. *The Rational Foundations of Social Institutions w Politics from Anarchy to Democracy,* pod redakcją I. Morrisa, J. Oppenheimera i K.E. Sołtana, Stanford University Press, Stanford CA, 2004.

Boyd, Robert, i Jeffrey Lorberbaum. 1987. *No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game*. „Nature" 327 (May 7): 58-59.

Boyd, Robert, i Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: The University of Chicago Press.

Cabrales, Antonio. 1993. *Stochastic Replicator Dynamics*. „Economics Working Paper" 54. Barcelona: Universitat Pompeu Fabra.

Coleman, James S. 1986. *Social Theory, Social Research, and a Theory of Action*. „American Journal of Sociology" 91: 1309-35.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge: Harvard University Press.

Dawkins, Richard. 1989. *The Selfish Gene*. Oxford: Oxford University Press.

Eggertsson, Thrainn. 1990. *Economic Behavior and Institutions*. Cambridge University Press.

Friedman, Daniel. 1991. *Evolutionary Games in Economics*. „Econometrica" 59: 637-66.

Fudenberg, Drew, i David K. Levine. 1998. *The Theory of Learning in Games*. Cambridge: MIT Press.

Fudenberg, Drew, i Jean Tirole. 1991. *Game Theory*. Cambridge, MA: MIT Press.

Gale, John, Kenneth Binmore, i Larry Samuelson. 1995. *Learning to be Imperfect: The Ultimatum Game.* „Games and Economic Behavior" 8 (January): 56-90.

Hechter, Michael. 1992. *The Insufficiency of Game Theory for the Resolution of Real-Life Collective Action Problems. „*Rationality and Society" 4: 33-40.

Hines, W.G.S. 1987. *Evolutionary Stable Strategies: A Review of Basic Theory*. „Theoretical Population Biology" 31: 195-272.

Lissowski Grzegorz i Piotr Swistak. 1998. *Wybór najlepszego uporządkowania społecznego: nowe zasady sprawiedliwości i normatywne wymiary wyboru*. „Studia Socjologiczne" 1998, 1 (148) pp.89-136.

Mailath, George J. 1992. *Introduction: Symposium on Evolutionary Game Theory*. „Journal of Economic Theory" 57: 259-77.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, John, i G. Price. 1973. *The Logic of Animal Conflict*. „Nature" 246: 15-18.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

Putnam, Robert D. 1993. *Making Democracy Work.* Princeton: Princeton University Press.

Rubinstein, Ariel. 1991. *Comments on the Interpretation of Game Theory*. „Econometrica" 59: 909-24.

Samuelson, Larry. 1993. *Recent Advances in Evolutionary Economics: Comments*. „Economics Letters" 42: 313-19.

Samuelson, Larry. 1998. *Evolutionary Games and Equlibrium Selection*. Cambridge: MIT Press.

Selten, Reinhard. 1983. *Evolutionary Stability in Extensive 2-Person Games*. „Mathematical Social Sciences 5: 269-363.

Selten, Reinhard. 1991. *Evolution, Learning, and Economic Behavior*. „Games and Economic Behavior" 3:3-24.

Simon, Herbert i Jonathan Schaeffer. 1992. *The Game of Chess*. In R.J.Aumann i S.Hart (eds.), „Handbook of Game Theory", volume 1. Elsevier Science Publishers B.V.

Sobel, Joel. 1993. *Evolutionary Stability and Efficiency*. „Economic Letters" 42 (2-3): 301-12.

Swistak, Piotr. 2003. *The Emergence and Stability of Social Capital*, APSA Meetings, Philadelphia, August 28-31.

van Damme, Eric. 1987. *Stability and Perfection of Nash Equilibria*. Berlin: Springer-Verlag.

Vega-Redondo, Fernando. 1996. *Evolution, Games, and Economic Behavior*. Oxford: Oxford University Press.

Vincent, Thomas L. i Joel S. Brown. 1988. *The Evolution of ESS Theory*. „Annual Review of Ecology and Systematics" 19: 423-43.

Warneryd, Karl. 1993. *Cheap Talk, Coordination, and Evolutionary Stability*. „Games and Economic Behavior" 5 (October): 532-46.

Weibull, Jorgen. 1995. *Evolutionary Game Theory*. Cambridge: MIT Press.

Young, Peyton. 1998. *Individual Strategy and Social Structure*. Princeton, N.J.: Princeton University Press.

Young, Peyton, i Dean Foster. 1991. *Cooperation in the Short and in the Long Run*. „Games and Economic Behavior" 3: 145-56.