
Munich Personal RePEc Archive

Evolution of altruism in the light of behavioral economics

Karbowski, Adam

Warsaw School of Economics

2011

Online at https://mpra.ub.uni-muenchen.de/69604/
MPRA Paper No. 69604, posted 21 Feb 2016 20:43 UTC

O kilku modelach samolubnego karania w ekonomii behawioralnej[1]

Adam Karbowski
Katedra Ekonomii II, SGH

Streszczenie

Artykuł omawia zjawisko samolubnego karania i jego wpływ na ewolucję zachowań altruistycznych. W pracy przedstawiono podstawowe modele teoretyczne wyjaśniające mechanizm samolubnego karania. Dyskusję nad praktycznymi implikacjami omawianej koncepcji wzbogacono licznymi przykładami. Na końcu postawiono nowe hipotezy badawcze wymagające weryfikacji empirycznej.

Słowa kluczowe: altruizm, karanie, kooperacja, ekonomia behawioralna, instytucje w społeczeństwie.

Wprowadzenie

Uwaga naukowców, którzy badają determinanty rozwoju gospodarczego krajów świata, skupia się coraz częściej na roli instytucji w życiu społecznym (MacKinnon et al., 2009). Instytucje pojmowane są tu bardzo szeroko jako zespół trwałych elementów ładu społecznego. Instytucje to mechanizmy regulujące życie w społeczeństwie poprzez dostarczanie powszechnie akceptowanych sposobów rozwiązywania konfliktów oraz problemów współpracy. Wśród przykładów instytucji wymienić należy: język, małżeństwo, rynek, religię, pieniądz, przedsiębiorstwo, zwyczaje i normy.

Zdaniem autora powstanie i rozwój wielu instytucji nie byłby możliwy bez międzyludzkiej kooperacji oraz altruizmu. Wydaje się, że zachowania prospołeczne stanowią fundament wielu systemów instytucjonalnych świata.

Niniejsza praca porusza zupełnie podstawową kwestię. Idzie tu o odpowiedź na pytanie, w jaki sposób kooperacja oraz altruizm powstają oraz ewoluują wraz z rozwojem społeczeństwa?

Okazuje się, że odpowiedź na tak postawione pytanie wymaga rzetelnego zbadania relacji pomiędzy zachowaniami altruistycznymi a zjawiskiem karania. Bogata literatura z kręgu ekonomii behawioralnej omawia i próbuje wyjaśnić fenomen **altruistycznego karania**. Mechanizm ten polega na ukaraniu osobnika postępującego w sposób samolubny, eksploatatorski lub nie zorientowany na współpracę w nadziei spowodowania zmiany jego

[1] Artykuł opublikowany w *Zeszytach Naukowych Kolegium Gospodarki Światowej*, 29, 236-249.

postępowania w przyszłości. Osobnik karzący nie odnosi przy tym żadnych realnych korzyści, karze ze względu na dobro innych.

O ile sam mechanizm altruistycznego karania jest względnie dobrze poznany i szeroko dyskutowany w literaturze (Fehr i Gächter, 2002), o tyle niezwykle ciekawe pozostaje zagadnienie **samolubnego karania**. Chodzi tu o odpowiedź na pytanie, czy osobnik, który dąży do osiągnięcia przede wszystkim dobra własnego kosztem dobra grupy, może podjąć się karania innych osobników, którzy w danej sytuacji współzależności społecznej zachowują się w sposób niekooperacyjny?

Jeśli odpowiemy na to pytanie twierdząco, zapytajmy, na ile takie zachowanie jest stabilne ewolucyjnie, jak osobnik karzący będzie postępował w kolejnych interakcjach społecznych i dlaczego w ogóle eksploatator podejmuje się karania innych eksploatatorów? Jakie są w końcu czynniki poznawcze i behawioralne, które mają wpływ na dynamikę procesu samolubnego karania?

Powyższe pytania będą stawiane w niniejszej pracy. Tezą autora pozostanie twierdzenie, że samolubne karanie pojawić się może w grupach zwartych o wyraźnie zarysowanej strukturze hierarchicznej. Ponadto, zdaniem autora, zrodzenie i stabilizowanie się samolubnego karania można częściowo wyjaśnić poprzez mechanizm **zachowań stadnych** i **brak pełnej informacji** w momencie dokonywania wyboru. Zachowania stadne zostaną zakwalifikowane tu jako czynnik behawioralny, niedoskonałość informacyjna natomiast jako czynnik poznawczy. Zarysowana tu koncepcja zostanie rozwinięta w dalszej części pracy.

Na początku chciałbym poczynić kilka uwag natury metodologicznej. W pracy będę posługiwał się formalnym językiem teorii gier (Rubinstein i Osborne, 1994), która w elegancki sposób pozwala modelować wiele sytuacji o charakterze współzależności społecznej. Definicje pojęć, do których będę się dalej odwoływał, przedstawiam poniżej.

Definicja 1. *Grą* nazywamy formalny zapis sytuacji decyzyjnej, w którym wyodrębnić można następujące elementy:

- skończony zbiór N (zbiór graczy, tj. zbiór podmiotów podejmujących decyzje),
- niepusty zbiór A_i dla $\forall i \in N$ (zbiór strategii dostępnych dla każdego gracza i),
- relację preferencji $\succeq_i$ na $A = \times_{j \in N} A_j$ dla $\forall i \in N$ (relację preferencji dla każdego gracza i na zbiorze możliwych wyników gry).

Definicja 2. *Równowagą Nasha* gry nazywamy profil strategii $a^* \in A$ spełniający dla każdego gracza $i \in N$ następujący warunek:
$\left(a_i^*, a_{-i}^*\right) \geq_i \left(a_i, a_{-i}^*\right)$ dla $\forall a_i \in A_i$.

Oznaczenie a_i^* możemy zinterpretować jako najlepszą odpowiedź gracza i – tego na strategie pozostałych graczy; analogicznie, a_{-i}^* jako najlepszą odpowiedź gracza nie i – tego (dowolnego gracza z wykluczeniem i – tego) na strategie pozostałych graczy.

Definicja 3. *Równowagą stabilną ewolucyjnie* (Maynard Smith, 1984) nazywamy każdą równowagę Nasha, dla której spełniony jest następujący warunek:
$\left(a_i^*, a_{-i}\right) >_i \left(a_i, a_{-i}\right)$ dla $\forall a_i \in A_i$.

Zauważmy, że strategia stabilna ewolucyjnie (tu: a_i^*) jest nie tylko najlepszą odpowiedzią gracza i – tego na optymalną strategię gracza nie i – tego (wynika to z równowagi Nasha, por. Definicja 2), ale jednocześnie jest najlepszą odpowiedzią gracza i – tego na dowolne odchylenie strategii gracza nie i – tego od jego strategii optymalnej (Maynarda Smitha warunek odporności strategii stabilnej ewolucyjnie na mutację strategii pozostałych graczy, por. Definicja 3).

Mówiąc inaczej, gracz i nie ma racjonalnej motywacji do zmiany strategii stabilnej ewolucyjnie. W konsekwencji w populacji składającej się z osobników przyjmujących strategię stabilną ewolucyjnie nie rozprzestrzeni się mutant przyjmujący strategię alternatywną.

Wielu psychologów, biologów i ekonomistów zadaje sobie pytanie, dlaczego altruizm w ogóle istnieje, dlaczego nie został odrzucony w ewolucji zachowań społecznych? Z punktu widzenia psychologii ewolucyjnej altruizm jest przecież niczym innym jak zachowaniem podnoszącym dostosowanie innego osobnika kosztem dostosowania własnego (za Hamiltonem, 1964).

Zagadnienie ewolucji altruizmu zostało w znacznym stopniu wyjaśnione dzięki dwóm poważnym propozycjom teoretycznym, tj. koncepcji **altruizmu zwrotnego** Roberta Triversa (1971), a także teorii **altruizmu krewniaczego** Williama Hamiltona (1964).

Młodsze prace (Fehr i Gächter, 2000; 2002; O' Gorman et al., 2005) zwróciły jednak uwagę, że żadna z dwóch wspomnianych koncepcji nie wyjaśnia kooperacji w sytuacji jednorazowych interakcji niespokrewnionych osobników. Próby znalezienia odpowiedzi na

tak postawione pytanie przyczyniły się do wprowadzenia pojęcia **kary** do teorii zachowań prospołecznych. Okazało się bowiem, że odpowiednio dotkliwa kara stabilizuje współpracę w wielu sytuacjach współzależności społecznej.

Początkowo uważano jednak, że karania podejmują się jedynie altruiści, którym zależy na przestrzeganiu norm społecznych. Poprzez ukaranie eksploatatora altruiści mieli uczyć go zachowań kooperacyjnych lub przynajmniej wykształcić u niego przekonanie, że łamanie zasad społecznych nieodłącznie wiąże się z karą. W warunkach naturalnych taką sankcją według Barclaya (2006) może być: krytyka, ostracyzm, groźba o charakterze fizycznym lub społecznym lub strata życiowego partnera. Z punktu widzenia psychologii ewolucyjnej ta ostatnia sankcja wydaje się wyjątkowo dotkliwa.

W warunkach laboratoryjnych karę modeluje się najczęściej jako pewien monetarny koszt, jaki ponosi zarówno karzący, jak i ukarany albo, prościej, jako wykluczenie eksploatatora z dalszych gier. Jak pokazują liczne badania (Fehr i Gächter, 2002; Barr, 2001; Ostrom, Walker i Gardner, 1992; Yamagishi, 1986):

- zawsze istnieje grupa uczestników eksperymentu, która decyduje się karać za zachowania nie zorientowane na współpracę,
- kara taka okazuje się skuteczna, tj. faktycznie podnosi poziom kooperacji w grze.

Najnowsze prace poświęcone altruizmowi (Nakamaru i Iwasa, 2006; Eldakar et al., 2007) zwracają jednak uwagę, że nie tylko altruiści podejmują się karania w sytuacjach współzależności społecznej. Eldakar i inni (2007) idą o krok dalej. Wykazują na podstawie przeprowadzonych symulacji, że altruizm i karanie wraz z rozwojem populacji stają się ujemnie skorelowane. Oznacza to, że altruizm rzeczywiście ewoluuje, ale podtrzymywany jest w czasie poprzez konkurencję eksploatatorów. Niektórzy z nich karzą bowiem dla własnego interesu za zachowania nie ukierunkowane na współpracę. Zjawisko takie nazwano **samolubnym karaniem**.

W następnej części niniejszego opracowania przedstawię dwa najważniejsze modele wyjaśniające mechanizm samolubnego karania. Będą to koncepcje Eldakara, Farrella i Wilsona[2] (2007), a także Nakamaru i Iwasy (2006). Tę sformalizowaną część pracy zamknę własną propozycją teoretyczną. Dalej przeprowadzona zostanie dyskusja wzbogacona przykładami samolubnego karania zaczerpniętymi zarówno ze świata ludzi, jak i zwierząt.

[2] Stąd w dalszej części pracy będę mówił o modelu EFW.

Modele samolubnego karania

Rozpocznijmy naszą analizę od propozycji Eldakara, Farrella i Wilsona (2007). Rozważmy zatem następującą wieloosobową grę ewolucyjną. Uczestnicy gry (populacji) różnią się między sobą skłonnością do altruizmu (zmienna A), a także skłonnością do karania (zmienna P). Zmienne przyjmują w populacji wartości od 0 do 1 w sposób losowy (obie zmienne mają rozkład jednostajny na odcinku $<0,1>$). Gra jest dwuetapowa. W pierwszym etapie każdy osobnik otrzymuje darowiznę o wartości E a następnie podejmuje decyzję, jaką część E przeznaczyć na cele wspólne grupy. Wartość każdej wpłaty do wspólnej kasy jest podwajana, dalej zaś jest równo dzielona pomiędzy członków populacji. Pozostała część darowizny pozostaje w posiadaniu osobnika (por. Wzór 1).

Wartość zmiennej A decyduje o wielkości wpłaty do wspólnej kasy. Altruiści wpłacają oczywiście odpowiednio więcej. Wysokość zysku pojedynczego osobnika po pierwszej rundzie gry zapisać możemy w następujący sposób:

$$E(1-A_i)+2E(\sum_{j=1}^{N} A_j)/N \qquad (1).$$

Pierwszy składnik sumy to pozostawiona przez gracza i - tego część darowizny, drugi składnik to korzyść z funduszu wspólnego. Zauważmy, że mamy tu do czynienia z konfliktem pomiędzy interesem indywidualnym a interesem grupowym. Jest to gra o strukturze dylematu więźnia (Kelley i Grzelak, 1972).

W etapie drugim gry uczestnicy mogą przeznaczyć część swoich zasobów na wykrycie (ang. *screening*) i ukaranie osobników postępujących w rundzie pierwszej w sposób skrajnie egoistyczny. Wraz ze wzrostem wartości inwestycji w karanie rośnie prawdopodobieństwo wykrycia i wykluczenia najmniej altruistycznego osobnika z dalszych gier. Indywidualna wartość inwestycji w karanie zależy od trzech czynników:

- surowości (P_i) karzącego jako jego stałej dyspozycji osobowościowej,
- skali eksploatacji w rundzie pierwszej, tj. wielkości efektu „pasażera na gapę” (Bornstein, 1992) wyrażonej jako: $(\sum_{j=1, j\neq i}^{N-1}(1-A_j))/(N-1)$,
- stopnia trudności (C) wykrycia najmniej altruistycznego osobnika.

Indywidualną wartość inwestycji w karanie zapisać możemy zatem w następujący sposób:

$$pun_i = P_i \frac{(\sum_{j=1, j\neq i}^{N-1}(1-A_j))}{N-1} C \quad (2).$$

Prawdopodobieństwo, że najmniej altruistyczny osobnik nie zostanie wykryty przez gracza i równe jest: $esc_i = 1 - P_i \frac{(\sum_{j=1, j\neq i}^{N-1}(1-A_j))}{N-1}$. Prawdopodobieństwo wykrycia[3] największego eksploatatora i usunięcia go z populacji wynosi: $rem = \left(1 - \prod_{i=1}^{n-1} esc_i\right) D$, gdzie D określa prawdopodobieństwo wykluczenia wykrytego „pasażera na gapę".

Omówiona powyżej gra była powtarzana wielokrotnie (badacze przeprowadzili w ten sposób symulację ewolucji zachowań prospołecznych). Okazało się, że początkowo skłonność do altruizmu i skłonność do karania pozostawały nieskorelowane. W miarę rozwoju populacji korelacja ta stała się jednak ujemna (por. Rysunek 1), to jest, eksploatatorzy odznaczali się istotnie wyższą skłonnością do karania niż altruiści.
W konsekwencji (por. Wzór 2) eksploatatorzy inwestowali proporcjonalnie więcej zasobów w karanie niż altruiści. Paradoksalnie okazało się zatem, że ci, którzy nierzadko oszukują i wyzyskują innych, stoją także na straży norm społecznych. Czynią tak jednak dla własnego interesu.

[Rysunek 1]

Próba interpretacji tego zaskakującego wyniku znajdzie się w następnej części niniejszego opracowania.

Zjawisko samolubnego karania zostało także zbadane przez naukowców japońskich. Nakamaru i Iwasa[4] (2006) proponują model bogatszy, ale też znacznie bardziej skomplikowany niż opis EFW.

Gracze zróżnicowani są tu ze względu na dwie **cechy**: skłonność do altruizmu oraz skłonność do karania. W ten sposób wyróżniamy cztery różne kombinacje (por. Rysunek 2)

[3] Przez któregokolwiek gracza.

[4] Stąd model ten nazwę w skrócie NI.

tych dyspozycji osobowościowych w populacji: czysty altruista (AN), altruista karzący (AP), czysty eksploatator (SN) i eksploatator karzący (SP).

Rozważamy jedynie gry dwuosobowe, w których w sposób losowy „spotykają się" osobniki o jednym z czterech powyższych profili osobowościowych. Jeśli altruista napotka eksploatatora, ponosi koszt c, eksploatator zaś odnosi korzyść równą b ($b \geq c > 0$). Jeśli dojdzie do interakcji altruisty z altruistą, obaj podnoszą swoje dostosowanie o $b - c$. W przypadku spotkania eksploatatora z eksploatatorem dostosowanie obu osobników pozostaje bez zmian.

[Rysunek 2]

Wprowadźmy teraz możliwość karania do naszego modelu. Jeśli osobnik karzący napotka eksploatatora, sam ponosi koszt nałożenia kary równy: $-q$, eksploatator ponosi zaś koszty związane z „napiętnowaniem": $-p$. Bilans wszystkich możliwych zysków i strat powstałych w wyniku interakcji społecznych w ujęciu NI przedstawia Rysunek 3.

[Rysunek 3]

Przetrwanie (przeżycie) osobnika zależy od sumy zdobytych przez niego punktów (skumulowanych wypłat) w kolejnych iteracjach (powtórzeniach gry). Prawdopodobieństwo śmierci osobnika maleje więc wraz ze wzrostem uzbieranej przez niego liczby punktów.

Komputerowe symulacje opisanego procesu dały następujący rezultat: strategia samolubnego karania „zwycięża" nad pozostałymi strategiami, tj. gwarantuje najwyższy poziom dostosowania. Co więcej, strategia samolubnego karania okazuje się w świetle modelu NI strategią stabilną ewolucyjnie (por. Definicja 3; Rysunek 4). Prawdopodobieństwo przeżycia karzącego eksploatatora dąży wraz z rozwojem populacji do jedności.

[Rysunek 4]

Zauważmy, że omówione koncepcje samolubnego karania odwołują się do stałych dyspozycji osobowościowych (skłonności do altruizmu oraz skłonności do karania). Posiadanie pewnej cechy warunkuje tu więc wystąpienie określonego zachowania.

Powyższe podejście charakterystyczne jest dla psychologii ewolucyjnej (Buss, 2001). Pozwala ono konstruować modele przejrzyste i eleganckie formalnie. Z punktu widzenia ekonomii behawioralnej są to jednak teorie statyczne, zachowanie zdeterminowane jest tu genetycznie, poza samym modelem. EFW i NI nie uwzględniają także wpływu kontekstu społecznego w momencie dokonywania wyboru przez osobnika, pozbawiono go w końcu możliwości uczenia się.

Moja propozycja teoretyczna nie jest konkurencyjna wobec koncepcji omówionych wcześniej. Stanowić ma jedynie próbę spojrzenia na samolubne karanie z nieco innej strony. Według mnie samolubne karanie może pojawić się w populacji jako swoisty „efekt uboczny" podejmowania decyzji w grupach hierarchicznych przy braku pełnej informacji w momencie dokonywania wyboru. Nie uzależniam tu tym samym samolubnego karania od posiadanego przez osobnika genotypu.

Rozważmy prosty model. Otóż, załóżmy, że jeden z osobników należących do grupy dopuścił się pewnego czynu, który wpływa na dobrostan innych. Grupa ma strukturę hierarchiczną. Załóżmy dalej, że indywidualna ocena czynu (θ) ma jednostajny rozkład prawdopodobieństwa na odcinku $<-1,1>$. Czynom szkodliwym dla grupy odpowiadają wartości ujemne. Czyny takie podlegają karze.

Ocena indywidualna czynu zależy od zbioru informacyjnego oceniającego (Ω). Obserwator (tu: gracz A) dokonuje indywidualnej oceny czynu. Ocena taka dokonywana jest niezwłocznie po jego zaobserwowaniu. Formalnie zapiszemy: $\theta^A(\Omega^A) \in U[-1,1]$, czyli indywidualna ocena zaobserwowanego przez A czynu (θ^A) przyjmuje wartość z rozkładu jednostajnego na odcinku $<-1,1>$. Ewaluacja ta zależy od informacji dostępnych A (Ω^A) w momencie podejmowania decyzji. Jeśli $\theta^A(\Omega^A) < 0$, obserwator skłonny jest ukarać osobnika za popełniony czyn. Załóżmy, że właśnie taki wariant ma miejsce w naszej analizie.

Ponieważ gracz A żyje w grupie, realizacja kary (na przykład ostracyzmu) ma charakter współzależny, tj. wymaga zgody wśród członków populacji. Zakładamy, że zgoda taka musi zostać „wypracowana" na wyższych szczeblach hierarchii. Dlatego też czyn zaobserwowany i oceniony przez A zostanie poddany ocenie gracza B, który zajmuje odpowiednio wyższe miejsce w hierarchii. Pomiędzy jej szczeblami panuje niedoskonała komunikacja, tj. θ^i nie stają się tu wspólną wiedzą.

Regułę decyzyjną gracza B zapisać możemy w następujący sposób: $E[\theta^A \mid \theta^A(\Omega^A) < 0] + \theta^B(\Omega^B) < 0$. B skłonny będzie ukarać za popełniony czyn, gdy suma jego ewaluacji i branej pod uwagę oceny A będzie mniejsza od zera. Ponieważ gracz A

opowiedział się za karą, w naszej formule obecne jest prawdopodobieństwo warunkowe: $\theta^A \mid \theta^A(\Omega^A) < 0$. Ponadto prawdziwa ewaluacja A jest graczowi B bezpośrednio niedostępna. Uwzględnia on zatem oczekiwaną wartość negatywnej oceny A, traktując ją tym samym jako swoisty **sygnał**.

Zauważmy, że $E[\theta^A \mid \theta^A(\Omega^A) < 0]$ wynosi: $-\frac{1}{2}$, co wynika z własności rozkładu jednostajnego. Teraz regułę decyzyjną gracza B możemy zapisać w prosty sposób: $\theta^B(\Omega^B) < \frac{1}{2}$.

Do realizacji kary (na przykład ostracyzmu) potrzebna jest jednak zgoda kolejnych członków populacji, którzy stoją wyżej w grupowej hierarchii. Wprowadźmy zatem do modelu gracza C, który ponownie ocenia czyn, kierując się własnym zasobem informacji, a także opiniami już wydanymi (sygnałami). Gracz C opowie się za karą, jeśli $\theta^C(\Omega^C) < \frac{1}{2} + \frac{1}{4}$[5]. *Per analogiam,* reguła decyzyjna gracza D to: $\theta^D(\Omega^D) < \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$.

Zauważmy, że prawa strona kolejnych reguł decyzyjnych tworzy ciąg liczbowy, który dąży w granicy do jedności. Oznacza to, że w **przypadku uogólnionym** reguła decyzyjna ostatniego w łańcuchu gracza i – tego będzie następująca: $\lim_{i \to \infty} \theta^i(\Omega^i) < 1$. Ponieważ warunek ten jest zawsze spełniony, gracz znajdujący się u szczytu hierarchii na pewno opowie się za karą.

Model pokazuje bardzo prosty proces formowania się **zachowań stadnych**. Zauważmy, że kolejni gracze są coraz mniej skłonni do zmiany decyzji podjętej przez poprzednika. Przyczyną występowania takiego zjawiska jest brak pełnej informacji w momencie dokonywania wyboru, a także sekwencyjne podejmowanie decyzji w grupie hierarchicznej.

Warto także podkreślić, że kontekst społeczny może tu górować nad dyspozycjami osobowościowymi podmiotu. Zauważmy, że decydent i – ty mógłby być ze względu na swój genotyp czystym eksploatatorem. Jednakże w naszym modelu zdecyduje się on karać. To sama sytuacja współzależności społecznej uczyni go eksploatatorem karzącym. Tym samym okazuje się, że środowisko społeczne może uczyć karania.

Przedstawione tu wnioski znajdują pewne potwierdzenie w badaniach Shinady, Yamagishi i Ohmury (2004). Naukowcy ci wykazali na podstawie przeprowadzonego eksperymentu, że kary są znacznie częstsze i dotkliwsze wobec członków własnej grupy niż

[5] Wyprowadzenie formuły analogiczne jak w przypadku gracza B.

wobec osobników spoza społeczności. Wydaje się więc, że sama sytuacja **grupowej współzależności** może sprzyjać karaniu nawet ze strony osobników genetycznie do tego nie predysponowanych. Badania Shinady, Yamagishi i Ohmury (2004) potwierdzają także znaczenie zwartości grupy (jej spójnej struktury) dla formowania się **tendencji do karania**.

O ile modele EFW i NI wyjaśniają pojawienie się i stabilizowanie samolubnego karania w oparciu o stałe dyspozycje osobowościowe, o tyle moja propozycja teoretyczna uzupełnia to podejście o wpływ kontekstu społecznego. Oba ujęcia wydają się jednak nieco przejaskrawione i jednostronne. Dlatego też sądzę, że fenomen samolubnego karania najlepiej wyjaśniałby model eklektyczny.

Dyskusja

Zachowanie eksploatatora, który karze innych za czyny nie zorientowane na współpracę wydaje się z pozoru irracjonalne. Oto osobnik skrajnie egoistyczny, któremu nie zależy na dobru grupy, ponosi koszty (energii, czasu) w celu zabezpieczenia dobrobytu innych. Po namyśle dojdziemy jednak do wniosku, że samolubne karanie jest niczym innym jak **inwestycją eksploatatora** w przyszłe większe zyski. Eksploatator, karząc w okresie t, „podbija stawkę" i oczekuje wyższej nagrody w czasie $t+1$.

Alternatywne wytłumaczenie pochodzenia samolubnego karania opiera się na **koncepcji podziału altruistycznych zadań w populacji** (Eldakar et al., 2007). Zarówno zachowanie kooperacyjne, jak i karanie są **dobrami publicznymi** (Fehr i Gächter, 2002). Ich dostarczenie wiąże się więc z indywidualnymi kosztami, które przekraczają wartość przyszłych korzyści z dobra dla osobnika. Dlatego, zdaniem Eldakara i innych (2007), ciężar dostarczenia obu dóbr publicznych jest rozłożony pomiędzy różnych członków populacji. Altruiści „zaopatrują" tym samym populację w skłonność do współpracy, eksploatatorzy zaś w skłonność do karania.

Okazuje się także, że eksploatatorzy są znacznie bardziej skuteczni niż altruiści w wykrywaniu i karaniu osobników nie ukierunkowanych na współpracę. Według Eldakara i innych (2007) oszuści posiadają pewne specyficzne umiejętności, które znacznie ułatwiają im rozpoznawanie innych oszustów. Do tych umiejętności zaliczyć trzeba: biegłość w posługiwaniu się strategiami oszustwa, zdolność kamuflażu, doświadczenie w walce.

Dlatego też eksploatatorzy konkurują ze sobą o nagrodę w postaci „żeru" na dobru wspólnym. W interesie eksploatatora leży podtrzymywanie współpracy w populacji, ponieważ zwiększa w ten sposób wartość wspólnego dobra. Z drugiej jednak strony

eksploatatorzy rywalizują ze sobą, aby zapewnić sobie jak największy udział w dobru wspólnym. Paradoksalnie okazuje się zatem, że altruizm w ewolucji zachowań społecznych zabezpieczany jest przez konkurujących ze sobą oszustów.

Zachowania o charakterze samolubnego karania można zauważyć zarówno w świecie ludzi, jak i zwierząt. Postępowanie mafii, która stwarza i zabezpiecza porządek społeczny w dzielnicach miast, w których władza publiczna jest słaba[6] jest niczym innym jak samolubnym karaniem. Organizacja przestępcza zapewnia ochronę wielu biznesom, stwarza ramy kooperacji potrzebne do wymiany handlowej, sprzyja rozwojowi gospodarczemu dzielnicy. Z drugiej zaś strony w konsekwentny sposób „żeruje" na dobru wspólnym.

Innym przykładem samolubnego karania może być postępowanie średniowiecznych rycerzy (Bisson, 1994). Rzetelne źródła historyczne podają, że wielu rycerzy (powoływanych przecież do utrzymywania ładu społecznego) było zwyczajnymi „zbirami", którzy rywalizowali ze sobą, aby czerpać korzyści z wyzysku biedoty miejskiej czy chłopstwa (za Eldakarem i innymi, 2007).

Wiele przykładów samolubnego karania można dostrzec także w świecie zwierząt. Wenseleers i inni (2005) omawiają specyficzne zachowanie strażników u os drzewnych (*Dolichovespula sylvestris*). Strażnicy ci mają za zadanie pilnować (i ewentualnie karać) robotnice, aby te pracowały i nie składały jaj. Ograniczenie liczby potomstwa do pewnego poziomu zwiększa korzyści wspólne dla całego roju. Okazało się jednak, że karzący strażnicy zachowują się w sposób skrajnie egoistyczny, ograniczając liczbę potomstwa innych robotnic, sami składają jaja.

Emery i Clayton (2001) omawiają ciekawy przykład samolubnego karania u sójek. Otóż, niektóre sójki dopuszczają się kradzieży pożywienia z gniazd innych sójek. Okazuje się ponadto, że sójki – złodziejki są znacznie bardziej agresywne niż te „uczciwe" i karzą złodziei za złupienie gniazda znacznie surowiej niż „sójki uczciwe".

Wada (1994) opisuje z kolei rywalizację u krabów (*Ilyoplax pusillus)*. Otóż, kraby zasiedlające pewne terytorium często współpracują ze sobą, aby każdy z nich mógł szybciej zbudować swoje gniazdo. Próby łamania współpracy są wówczas bezwzględnie karane. Po wybudowaniu gniazd zwierzęta te okazują się jednak skrajnie egoistyczne, potrafią błotem i patykami ogrodzić wyjście z gniazda u sąsiada. W ten sposób zwiększają obszar swojego terytorium a sąsiada niejako skazują na śmierć z braku pożywienia. Początkowa współpraca przy budowie gniazd jest zatem niczym innym jak kooperacją karzących samolubów.

[6] Na przykład niektóre dzielnice Nowego Jorku czy Detroit w Stanach Zjednoczonych Ameryki Północnej.

Podsumowanie i dalsze pytania badawcze

Niniejsza praca poświęcona była zagadnieniu samolubnego karania. Pojęcie to zostało wprowadzone bardzo niedawno[7] do teorii zachowań prospołecznych. Wydaje się jednak, że samolubne karanie może być zachowaniem znacznie bardziej rozpowszechnionym w przyrodzie i o większym znaczeniu dla ewolucji altruizmu niż karanie altruistyczne.

Wciąż jednak bardzo wiele aspektów samolubnego karania wymaga wyjaśnienia. Istniejące modele są jednostronne, opierają się ponadto na wielu upraszczających założeniach.

Na przykład model EFW w sposób niejawny zakłada, że skłonność do altruizmu i skłonność do karania są zupełnie oddzielnymi cechami, a więc dyspozycjami osobowościowymi związanymi z różnymi genami. Stąd w pierwszej populacji modelu cechy A i P pozostają nieskorelowane. Ponadto wraz ze wzrostem liczebności grupy kara w EFW staje się nieefektywna, tj. nie podnosi poziomu kooperacji w grze. Wnioski z EFW odnoszą się więc jedynie do grup zwartych, w których koszty koordynacji nie są zbyt duże.

W końcu należy podkreślić, że EFW i NI są jedynie propozycjami teoretycznymi, nie były one weryfikowane empirycznie. Oba modele powstały na bazie symulacji przeprowadzonych w pakiecie matematycznym.

Pierwszym zadaniem w dalszych studiach nad zagadnieniem samolubnego karania byłaby więc próba przeprowadzenia eksperymentu według scenariusza gry wykorzystanej w EFW. Przed eksperymentem każdy z jego uczestników powinien zostać przebadany kwestionariuszem pozwalającym na określenie nasilenia jego skłonności do zachowań altruistycznych. Myślę, że wystarczającym narzędziem byłby tu inwentarz NEO – FFI, w którym wysoki wynik na skali ugodowości świadczyłby o skłonności do altruizmu[8]. Dalej należałoby zbadać związek pomiędzy wynikiem osoby w teście przed badaniem a, na przykład, liczbą decyzji o ukaraniu w wieloetapowej grze. Czy rzeczywiście okazałoby się, że ujawnia się ujemna korelacja pomiędzy skłonnością do altruizmu a karaniem? Jeśli tak, jaka byłaby siła tego związku? Czy zgodna z oczekiwaniami modelu EFW?

Drugim, niezmiernie ważnym zagadnieniem jest odpowiedź na pytanie: jak długo eksploatator będzie ponosił koszty związane z karaniem, zanim „skapitalizuje" swoją

[7] Pierwsze zaczęły powstawać w 2004 roku.

[8] Odpowiednio wynik niski świadczyłby o skłonności do zachowań samolubnych.

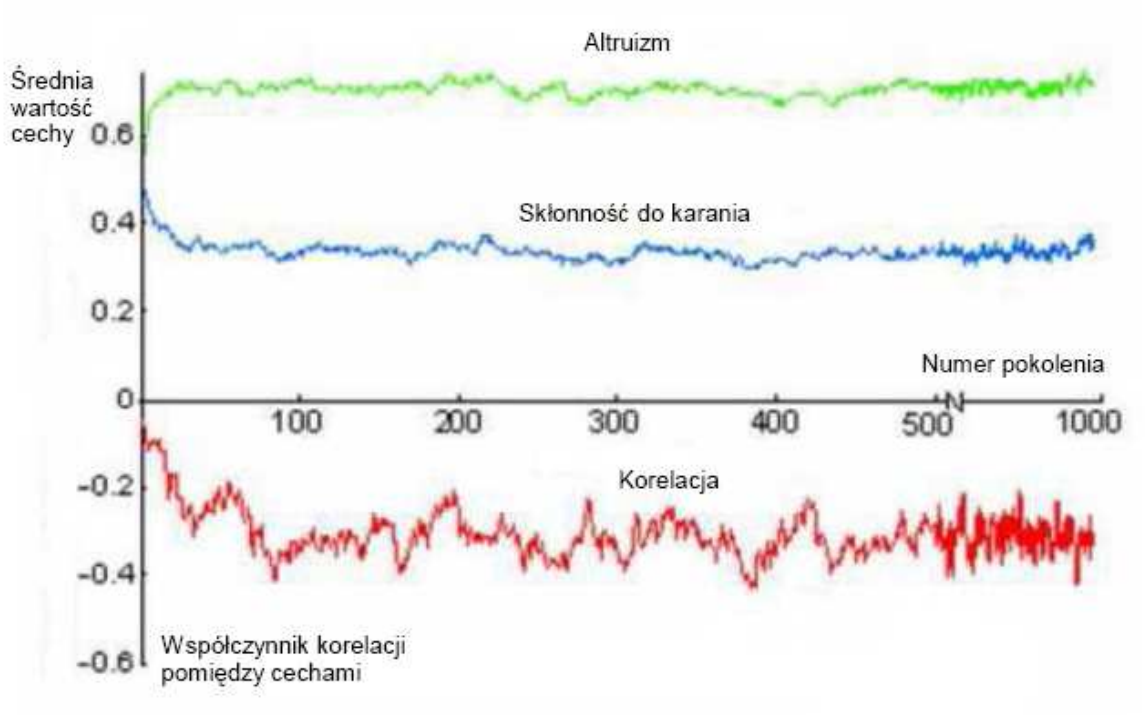
nagrodę? Jak bardzo eksploatator jest cierpliwy, jak długo jest w stanie odraczać moment uzyskania zysku?

Jasne jest, że im dłużej eksploatatorzy będą powstrzymywać się od oszustwa, tym większa „stawka" będzie do wygrania. Zbyt długie zwlekanie wiąże się jednak z rosnącym ryzykiem uprzedzenia przez innego eksploatatora. Warte zbadania wydaje się zatem zagadnienie relacji pomiędzy **tempem dyskontowania** a preferencjami wobec ryzyka w grupie eksploatatorów.

Trzecim obszarem dalszych badań byłaby próba włączenia do modeli samolubnego karania asymetrii w zakresie władzy (Clutton – Brock i Parker, 1995; Giraldeau i Caraco, 2000; Kim, 2006; Monnin i Ratnieks, 2001). Modele EFW i NI zakładają równość statusu społecznego członków populacji. Jest to jednak założenie nierealistyczne. Zarówno w świecie ludzi, jak i zwierząt niektórzy członkowie grupy mają znacznie większe możliwości w zakresie karania niż pozostała część populacji.

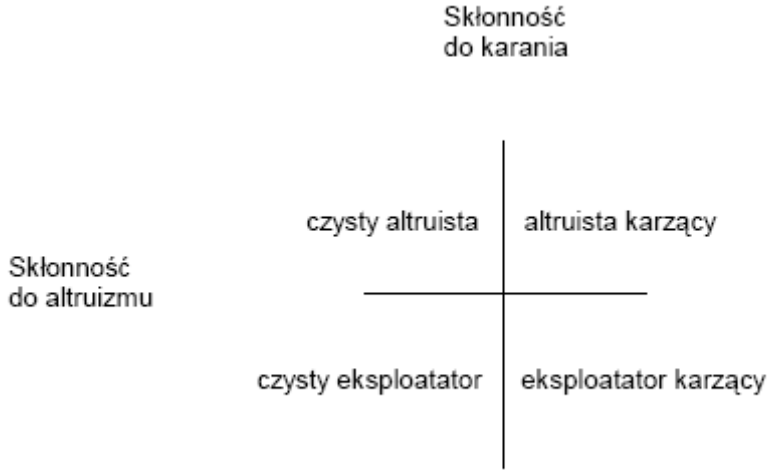
Chociaż koncepcja samolubnego karania wciąż niesie ze sobą wiele pytań i wątpliwości, pozwoliła ona jednak spojrzeć na ewolucję altruizmu z zupełnie innej strony niż dotychczas. Uznane propozycje teoretyczne (Trivers, 1971; Hamilton, 1964) zakładają, że to sami altruiści zabezpieczają swoje przetrwanie. Osiągają to poprzez odpowiednio częstsze udzielanie pomocy innym altruistom. Koncepcja samolubnego karania nie nakłada podobnych warunków ograniczających na grupę altruistów. Przewrotnie stwierdza natomiast, że altruiści rozwijają się i mają potomstwo, ponieważ w pewnej mierze jest to sytuacja pożądana przez egoistów. Tak nowe spojrzenie może być inspirujące dla badaczy zajmujących się problematyką zachowań prospołecznych i, moim zdaniem, przyczyni się do wydania wielu oryginalnych i interdyscyplinarnych prac naukowych.

Rysunek 1: Wyniki symulacji przeprowadzonej na podstawie EFW.



Źródło: Eldakar, O.T., Farrell, D.L. i Wilson, D.S. (2007). Selfish punishment: Altruism can be maintained by competition among cheaters. *Journal of Theoretical Biology*, w druku.

Rysunek 2: Możliwe kombinacje dyspozycji osobowościowych graczy w modelu NI.



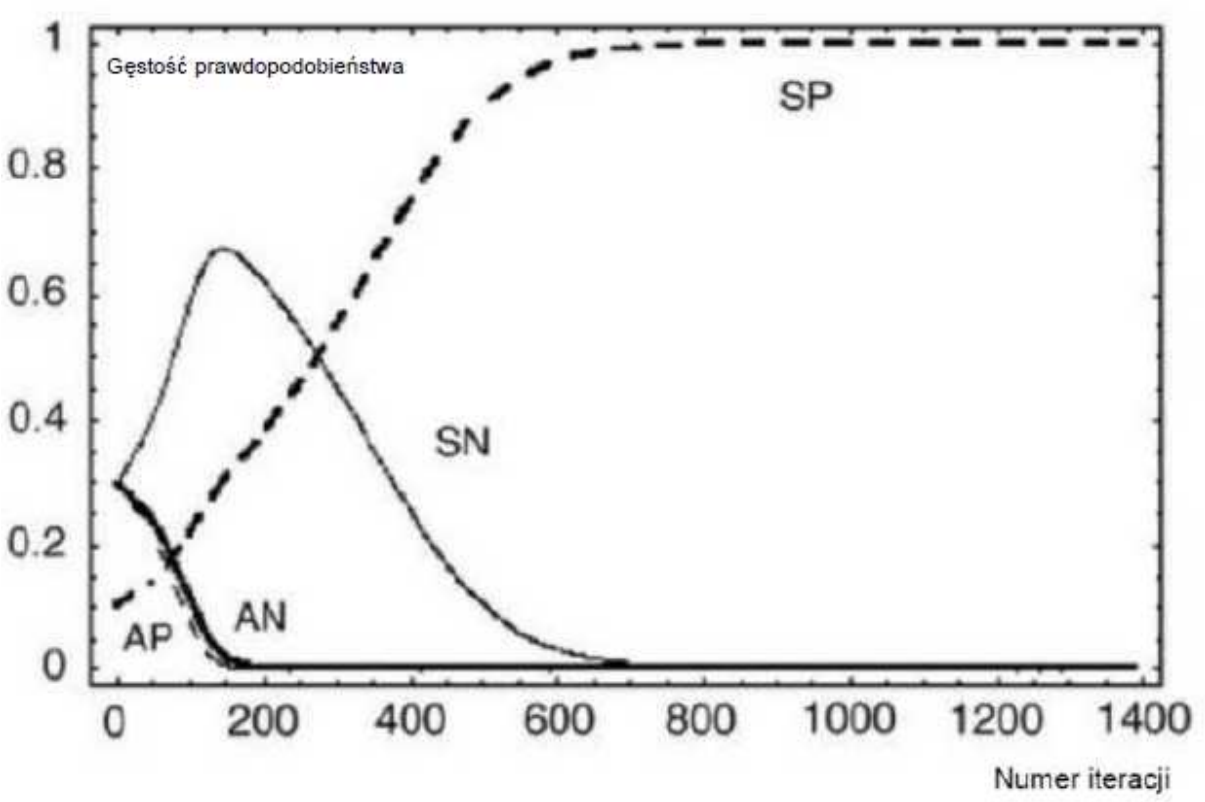
Źródło: Opracowanie własne.

Rysunek 3: Możliwe wypłaty w dwuosobowej grze NI.

	Gracz 2			
Gracz 1	AP	AN	SP	SN
AP	$b-c$	$b-c$	$-c-q$	$-c-q$
AN	$b-c$	$b-c$	$-c$	$-c$
SP	$b-p$	b	$-q-p$	$-q$
SN	$b-p$	b	$-p$	0

Źródło: Nakamaru, M. i Iwasa, Y. (2006). The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology,* 240, 475-488.

Rysunek 4: Prawdopodobieństwo przeżycia osobnika o jednym z czterech możliwych profili osobowościowych w modelu NI w różnych momentach czasu.



Źródło: Nakamaru, M. i Iwasa, Y. (2006). The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology,* 240, 475-488.

Literatura

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*, 325 – 344.

Barr, A. (2001). Social dilemmas and shame – based sanctions. Working Paper WPS/2001/11, University of Oxford, UK.

Bisson, T. (1994). The feudal revolution. *Past and Present, 142*, 6 – 42.

Bornstein, G. (1992). The free rider problem in intergroup conflicts over step-level and continuous public goods. *Journal of Personality and Social Psychology*, *62*, 597 – 606.

Buss, D. (2001). *Psychologia ewolucyjna*, Gdańsk, GWP.

Clutton – Brock, T. i Parker, G. (1995). Punishment in animal societies. *Nature, 373*, 209 – 216.

Eldakar, O.T., Farrell, D.L. i Wilson, D.S. (2007). Selfish punishment: Altruism can be maintained by competition among cheaters. *Journal of Theoretical Biology*, w druku.

Emery, N. i Clayton, N. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature, 414*, 443 – 446.

Fehr, E. i Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137-140.

Fehr, E. i Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*, 980-994.

Giraldeau, L. i Caraco, T. (2000). *Social Foraging Theory,* Princeton University Press, Princeton, NJ.

Hamilton, W.D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, *7*, 1-16 i 17-52.

Kelley, H. i Grzelak, J. (1972). Conflict between individual and common interest in an n-person relationship. *Journal of Personality and Social Psychology, 21*, 190 – 197.

Kim, S. (2006). For whom do we reciprocate? The effects of dominance relationships on the use of incentives in collective action. Working Paper.

MacKinnon, D., Cumbers, A., Pike, A., Birch, K. i McMaster, R. (2009). Evolution in Economic Geography: Institutions, Political Economy, and Adaptation. *Economic Geography, 85,* 129-150.

Maynard Smith, J. (1984). Game theory and the evolution of behaviour. *Behavioral and Brain Sciences*, *7*, 95-125.

Monnin, T. i Ratnieks, F. (2001). Policing in queenless ponerine ants. *Behav Ecol, 50*,

97 – 108.

Nakamaru, M. i Iwasa, Y. (2006). The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology, 240*, 475-488.

Nęcka, E., Orzechowski, J. i Szymura, B. (2006). *Psychologia poznawcza*, Warszawa, PWN.

O' Gorman, R., Wilson, D.S. i Miller, R.R. (2005). Altruistic punishing and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evolution and Human Behaviour, 26*, 375-387.

Ostrom, E., Walker, J. i Gardner, R. (1992). Convenants with and without a sword: Self-governance is possible. *American Political Science Review, 86*, 404 – 417.

Rubinstein, A. i Osborne, M. (1994). *A Course in Game Theory,* Cambridge, MIT Press.

Shinada, M., Yamagishi, T. i Ohmura, Y. (2004). False friends are worse than bitter enemies: "altruistic" punishment of in-group members. *Evolution and Human Behavior, 25*, 379 – 393.

Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 35-57.

Wada, K. (1994). Earthen structures built by *Ilyoplax pusillus*. Ethology, *96*, 270 – 282.

Wenseleers, T., Tofilski, A. i Ratnieks, F. (2005). Queen and worker policing in the tree wasp *Dolichovespula silvestri*. *Behav Ecol Sociobiol, 58*, 80 – 86.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*, 110 – 116.

Summary in English

Evolution of altruism in the light of behavioral economics

The article discusses the phenomenon of selfish punishment and its impact on the evolution of altruistic behavior. The paper presents the basic theoretical models explaining the mechanism of selfish punishment. Discussion of the practical implications of this concept is enriched with numerous examples. At the end of the paper new research hypotheses were formulated and all of them require empirical verification.

Keywords: altruism, punishment, cooperation, behavioral economics, institutions in society